Whole genome analysis of copy number variation in a case control study of recurrent depressive disorder

Rucker, James

Awarding institution:
King's College London

WHOLE GENOME ANALYSIS OF COPY NUMBER VARIATION IN A CASE CONTROL STUDY OF RECURRENT DEPRESSIVE DISORDER

James J.H. Rucker

Submitted to the University of London for the degree of Doctor of Philosophy

March 2012

Social, Genetic and Developmental Psychiatry Centre
The Institute of Psychiatry
King's College London

Abstract

Rare copy number variants (CNV), defined as deletions and duplications of genetic material over 1,000 base pairs in length, have become the focus of considerable interest in psychiatric disorders, where a proportion of individuals harbour rare and de novo events not usually seen in controls. We have performed a genome wide association study of copy number variation in 3,106 cases of recurrent depressive disorder, 1,731 controls screened for a lifetime absence of psychiatric disorder, and 5,619 population controls from phase 2 of the Wellcome Trust Case Control Consortium. Analysing our data with the PennCNV method, we found an enrichment of rare deletion CNVs in our case cohort, especially when compared to our screened control cohort. This finding was supported by further analysis with the iPattern method, but not by the QuantiSNP method. We followed up a selection of cases and controls with a comparative genomic hybridisation (CGH) array focussed on the region 22q11.2, which is a neuro-gene rich region of the human genome under current active evolutionary selection and resident to a deletion syndrome which commonly manifests with psychiatric disorders. We found no significant differences in CNV burden between our case and control cohorts. Finally we ran association analyses with our CNV call sets, including a high quality intersected call set derived from all three methods, against various phenotypes obtained from a combined database of all studies that contributed samples to this GWAS. We found no associations that survived Bonferroni correction for multiple testing.

Statement of Work

The genome wide association study in recurrent depressive disorder is a large, multi-centre study coordinated at the Social, Genetic and Developmental Psychiatry Centre at the Institute of Psychiatry. The principle investigators (PIs) for the clinical and DNA collections were Professor Peter McGuffin and Professor Anne Farmer and for the genome wide association study their co-PIs were Dr Gerome Breen and Professors Cathryn Lewis and Ian Craig. The study is composed of samples derived from the Depression Case Control study (DeCC), the Depression Network study (DeNT), the Genome Based Therapeutic Drugs for Depression study (GENDEP) (on which Dr Katherine Aitchison was also co-PI) and the Bipolar Case Control study (BaCC) for control samples only. We also used data from control samples from phase 2 of the Wellcome Trust Case Control Consortium (the national blood service cohort and the 1958 birth cohort).

Sample genotyping was performed by the Centre Nationale De Genotypage (CNG) in Evry, Paris (lead, Mark Lathrop). Cleaning of individual samples and SNP genotypes for the common CNV analysis was performed by Mandy Ng and Cathryn Lewis.

CNV calling using PennCNV and QuantiSNP was performed by James Rucker. iPattern calling was performed in Toronto by Dalila Pinto, Zhuozhi Wang and James Rucker. Sample QC and analysis was performed by James Rucker. Intersection of calls from three methods was performed by James Rucker and

Bhooma Thiruvahindrapuram. All other bioinformatics, plotting and analysis was performed by James Rucker. Scripting assistance in perl, R and STATA was given by Inti Pedroso, David To, Katherine Tansey and Rudolf Uher.

Preparation of samples and analysis of data from Agilent comparative genomic hybridisation arrays was carried out by James Rucker. Array design and processing was performed by Oxford Genome Technologies (Oxford, UK).

Phenotype data was collected as part of each individual study within the GWAS. Additional phenotype and neuroimaging data for a subset of participants followed up for other studies, but used in this thesis, was collected by James Rucker and James Cole.

Unless otherwise stated, all genomic coordinates stated in this work are based on build 18 of the human genome reference sequence (March, 2006. NCBI36/hg18).

Acknowledgements

I would like to dedicate this thesis to my parents, whose tireless devotion to the loving and stable upbringing I was lucky enough to experience as a child has given me the solid foundation upon which I have been able to build my relationships, personal interests and career, culminating in this work. I would also like to dedicate this thesis to my brothers Nick and Si and my sister-in-law Minsi for their unflinching love and support when times have been tough, and through all the good times too.

And to my friends, whose support has been essential throughout the years of work this thesis is based on, I also offer my warm and heartfelt thanks, particularly Ro Podolczuk, Samin Saeed, Harry Blatch, Noel Collins, Ben Siegal, Justin Wakefield, Elliot Peel, Jo and Ray Anderson, Dave Llewellyn, Will Tatam, Jack Franks, Liz Searle, Helen Wiliams, Katie Groves and Andy Rushton.

In no particular order I would like to thank my colleagues at work, many of whom have also become firm friends; Margarita Rivera-Sanchez, Sarah Cohen-Woods, Shazza Al-Sabban, Katherine Tansey, Inti Pedroso, Katherine Naverette, Cerisse Gunasinghe, Joanna Gray, Ursula Paredes, Jonathan Huntley and Toby Winton-Brown.

Last, but by no means least, I would also like to thank my supervisors, Professor Peter McGuffin and Dr Gerome Breen, without whom I would have steered a course beset by theoretical misconceptions and blind analytical alleys.

This PhD project was funded by the Wellcome Trust, and I am most grateful to them for providing the generous and flexible funding that paid my salary, allowed me to broaden my horizons and ultimately made this research a reality.

Contents

List of Tables

List of Figures

Chapter 1. Introduction

"Today will be clouded with self-doubt followed this evening with a blanket of denial . Tomorrow will find a clearing of anxiety followed by a sunny disposition."

1.1 Major Depression

The prevalence of psychiatric disorders in the general population is surprisingly high (Kessler et al., 1994). Within high income countries the World Health Organisation estimates that unipolar depression is the leading cause of Disability Adjusted Life Years, defined as the number of years lost due to ill-health, disability or early death(Mathers, Fat, World Health Organization, & Boerma, 2008). The World Health Organisation estimates that the worldwide prevalence of unipolar depression is about 121 million people, that it is amongst the leading causes of disability worldwide, that it can be reliably diagnosed in primary care but that fewer than a quarter of sufferers have access to effective treatments (World Health Organisation, n.d.).

A depressive episode is defined within the World Health Organisation's International Classification of Diseases, 10th edition, (ICD-10) as a pervasive lowering of mood, reduction of energy and decrease in activity with a reduced capacity for enjoyment, interest and concentration lasting for at least two weeks(World Health Organisation, 1993; 1993). Biological symptoms including sleep and appetite disturbance are commonly seen. The above symptoms are almost always associated with a lowering of self-esteem and self-confidence with ideas of guilt and unworthiness. The depressed mood is pervasive and varies little from day to day. According to the number of symptoms present and the degree of disability conferred, a depressive episode may be classed as mild, moderate, severe or severe with psychotic symptoms (World Health Organisation, 1992). Psychotic symptoms, if present, must be congruent with

mood. That is, the nature of the delusions or hallucinations is in keeping with the observed affective state.

Perhaps because the symptoms of depressive disorder overlap with normal human responses to stress or trauma, the prevalence of the disease has been widely estimated. The large Epidemiological Catchment Area study in the United States reported a lifetime population prevalence of depressive disorder at 4.4%(Weissman et al., 1988), whereas the National Comorbidity Survey estimated the prevalence at 17.1% (Kessler et al., 1994). Prevalence estimates are highly dependent on the severity cut-off and definitions employed by different research groups. However one striking feature throughout prevalence estimates for depression is that women demonstrate an approximately two-fold increased prevalence when compared to men(Kessler et al., 1994).

A minority of patients with depressive disorder will also suffer distinct manic episodes consisting of increased mood, increased activity and often reckless behaviour. Whilst the two states are often conceptualised to be two extremes of one continuum, if a patient with a history of depressive episodes is diagnosed with a manic episode they are consequently diagnosed with bipolar affective disorder, which is considered to be a distinct nosological entity within both the Diagnostic and Statistical Manual, 4th edition(American Psychiatric Association, 1994) (DSM-IV) and ICD-10.

The treatment of depressive disorders depends largely on the severity. Mild cases are often treated conservatively (so-called 'watchful waiting'), as many cases will resolve spontaneously. Moderate and severe forms of depressive

illness are treated either with brief psychotherapeutic interventions such as cognitive behavioural therapy, antidepressant medication, or both. Most treatment is carried out in primary care settings in the UK however treatment resistant and comorbid depressive disorder is often referred to specialist psychiatric services. Hospitalisation is usually reserved for high risk cases with poor social support. Concomitant psychotic symptoms may also be treated with antipsychotic medication. In severe cases, where threat to life is present from dehydration or malnutrition, electroconvulsive therapy is sometimes used.

1.2 Recurrent Depression

Recurrent depression is defined as two or more depressive episodes, separated by a period of at least two months(World Health Organisation, 1992). The ICD-10 distinguishes single depressive episodes and recurrent depression by separating them into distinct sub-sections of the same category. Prior research suggests that many individuals who suffer one depressive episode will go on to develop another, with those who have suffered three or more episodes particularly vulnerable to further episodes(Keller, Shapiro, Lavori, & Wolfe, 1982; Zis & Goodwin, 1979). The data suggest that unipolar depression, like bipolar disorder, is better conceptualised as a chronic, relapsing/remitting disorder with increasing risk for recurrence with each successive episode (Kessing, Hansen, Andersen, & Angst, 2004). In this sense and in combination with the high prevalence rate, the magnitude of the disabling effect of unipolar depressive disorder can be better conceptualised, and it also neatly encapsulates

why research into this disorder, where treatment in many cases does not lead to full remission, is warranted (Fava et al., 1997).

1.3 The Genetics of Depressive Disorder

1.3.1 Family, Adoption and Twin Studies

The familial aggregation of mood disorders has been noted for over 100 years(M. Tsuang & Faraone, 1990). Within formal family studies, probands who suffer from depressive disorder are compared with demographically matched controls, and the prevalence of major depression in first degree (or other) relatives is ascertained. In a meta-analysis of five family studies(Gershon et al., 1982; Maier et al., 1993; M. T. Tsuang, Winokur, & Crowe, 1980; Weissman et al., 1984; 1993) undertaken in major depression, Sullivan et al.(Sullivan, Neale, & Kendler, 2000) reported that the first degree relatives of probands with depressive disorder were 2.83 fold more likely to suffer from the disorder than controls.

Adoption studies feature the offspring of one set of parents who happen to be reared from early life by unrelated parents. Of three major adoption studies, two(Knorring, Cloninger, Bohman, & Sigvardsson, 1983; Wender et al., 1986) support the idea of genetic factors being significantly involved in the aetiology of major depression, whilst one other(Cadoret, O'Gorman, Heywood, & Troughton, 1985) did not (although in a re-analysis of combined groups within this study, Sullivan et al.(Sullivan et al., 2000) did find a significant difference).

Twin studies feature and contrast monozygotic and dizygotic twins. Because monozygotic twins broadly share all their genes in common, and dizygotic twins

half their genes in common, whilst both sharing many of the same environmental features, twin studies allow estimates of heritability to a disorder to be calculated. Sullivan et al., in their meta-analysis of five studies (Bierut et al., 1999; Kendler & Prescott, 1999; Kendler, Pedersen, Neale, & Mathé, 1995; Lyons et al., 1998; McGuffin, Katz, Watkins, & Rutherford, 1996), estimated the heritability of depressive disorder to be 37% (Sullivan et al., 2000). However this figure increases to approximately 70% when samples with recurrent and severe forms of the disease are analysed (Kendler, Neale, Kessler, Heath, & Eaves, 1992; McGuffin et al., 1996).

Similarly, whilst the relative risk to siblings of probands with depressive disorder is about 3, this increases to over 9 when much stricter definitions of depression and health are used (Farmer et al., 2000; I. Jones, Kent, & Craddock, 2002). A graduation of risk is seen as genetic distance from the proband increases, with monozygotic co-twins having a risk of about 50% of also suffering from depression(Craddock & Forty, 2006). However the aetiological and clinical heterogeneity of depressive illness probably precludes estimates that are generalizable to a large proportion of sufferers. Put another way, the spectrum of aetiology of depressive episodes seen in the clinic is very wide. Individuals with early onset, severe forms of the disorder may be particularly at risk of manic episodes, and may carry a more heritable form of the disease(Kendler, Gardner, & Prescott, 1999). Those with pre-existing anxious and dysthymic (neurotic) personalities may have genetically influenced personality factors or have suffered adverse experiences in childhood (or both)(Benjamin, Ebstein, & Belmaker, 2002; Kendler et al., 2000). Depression in

later life may be more related to other co-occurring pathological processes such as endothelial dysfunction(Potter et al., 2007). Combining putative subgroups together in large studies of depression is practical, but carries with it problems associated with clinical heterogeneity.

1.3.2 Linkage and Genome Wide Association (GWA) Studies

Within the context of family, twin and adoption studies, several large whole-genome linkage studies on samples with recurrent depression were completed within the last 10 years, the Utah sample, the Depression Network (DeNt) study and the Genetics of Recurrent Early-Onset Depression (GenRED) study(Camp et al., 2005; Holmans et al., 2004; McGuffin et al., 2005). All three support linkage in the 15q region, although the relative risk is modest, not an unusual result for a study in a complex disorder. Two recent papers have demonstrated a significant linkage association on chromosome 3p25-26(Breen et al., 2011; Pergadia et al., 2011).

Genome-wide association studies (GWAS) are more powerful than linkage studies at detecting loci of small effect, and have become practicable with the advent of micro-array technology. Many common traits and disorders have had notable successes in identifying novel susceptibility genes, for example type 2 diabetes(Zeggini, 2007) and rheumatoid arthritis(Thomson et al., 2007). However even with GWAS there have been inconsistent findings in depressive disorder, with no replicated associations thus far forthcoming(Lewis et al., 2010; Muglia et al., 2010; Rietschel et al., 2010; Shi et al., 2011; Sullivan et al., 2000; Terracciano et al., 2010; Wray et al., 2012). Sullivan et al. implicated the gene

PCLO (piccolo), involved in monoaminergic transmission(Sullivan et al., 2009). Lewis et al. implicated the gene BICC1, although this failed confirmation in a replication analysis(Lewis et al., 2010). A recent meta-analysis by Shyn et al.(Shyn et al., 2009) included data from three of these studies(Muglia et al., 2010; Shi et al., 2011; Sullivan et al., 2009). They reported three intronic SNPs in ATP6V1B2, SP4, and GRM7 as most associated, but with p-values just short of genome wide significance.

Overall there is a relative paucity of understanding of the genetic predispositions to depressive disorder, although copy number variants, which can also be detected using micro-arrays, have not yet been comprehensively studied.

1.4 DNA Microarray Technology

A draft sequence of the human genome was first published at the turn of the millennium(Lander et al., 2001; Venter et al., 2001). At approximately the same time genomic microarrays, which essentially represent miniaturisations of the Southern blotting technique(Southern, 1975; 2006), started to become available(Schena, Shalon, Davis, & Brown, 1995). They allowed the interrogation of the genome at thousands of unique oligonucleotide sequences of known position in the genome reference on a single slide. This allowed massively parallel genomic interrogation studies to be carried out relatively cheaply on large numbers of samples. Since the advent of the reference human genome sequence advances have been rapid(Lander et al., 2001). Hundreds of thousands or millions of unique genomic probes are attached to a single modern microarray.

Each probe is attached at a known location on the chip and highly sensitive fluorescence analysers used to detect the degree of fluorescence at each location. This data can then be used, with appropriate normalisation and transformation, to infer the relative abundance of the sample DNA being tested. This fluorescence data is aggregated and normalised between redundant probes, and a continuous fluorescence value is then used either to arbitrate between bi-allelic SNP calls, or make a relative estimation of copy number from the relative fluorescence intensity compared to an internally derived canonical reference value.

Microarrays using single nucleotide polymorphism (SNP) markers capitalise on the fact that SNPs represent unique, known, biallelic markers. Each allele is tagged with a different flurophore, and thus the data produced for each SNP marker is essentially two-dimensional and additional information can be deduced from the relative intensity of the two flurophores. Modern arrays also include monoallelic probes to represent areas of the genome where SNPs are scarce, such as areas rich in repeat sequences. Such markers, by definition, produce a 1 dimensional spectrum of fluorescence data, however this information can be used to infer the relative abundance of that sequence in the DNA sample being studied, similar (but not identical) to a comparative genomic hybridisation (CGH) experiment. Modern microarrays from the main vendors (Affymetrix and Illumina) cover the majority of the human genome, and are now well established technologies. The International Standard Cytogenomic Array Consortium recently recommended microarrays as the first-line test for paediatric developmental disorders (Miller et al., 2010).

1.5 Copy Number Variants

The human genome is subject to many forms of variation. Large deletions and duplications of genetic material may be visualised with the light microscope on a stained karyotype spread, however the lower limit of resolution is approximately 3MB. On the basis of cytogenetic evidence, it was thought that deletion and duplication events were rare phenomena, explaining only a handful of mainly sporadic, infrequently occurring genetic diseases(Perry et al., 2008). The new wave of genetic data from GWAS showed that deletion and duplication events in fact occurred frequently in the normal population, and the extent of this was a surprise to many geneticists (Freeman et al., 2006; 2006; Iafrate et al., 2004; McCarroll et al., 2006; Sebat et al., 2004). Deletions and duplications of genetic material, which is now termed copy number variation, represents a large and dynamic form of genetic variation and the formation of gametes with deletions or, perhaps more importantly, duplications is likely to be an essential driver of species evolution(Locke et al., 2003; J. Lupski, 2007). Copy number variants are now known to be associated with a wide range of disorders, from obesity(Bochukova et al., 2009) to autism(Sebat et al., 2007). Within the field of psychiatric genetics this new realm of variation has rapidly become of considerable interest.

A copy number variant, or CNV, is rather arbitrarily defined as a deletion or duplication of genomic material of more than 1,000 base pairs(Feuk, Carson, & Scherer, 2006; Iafrate et al., 2004; Sebat et al., 2004). Deletion CNVs delete one copy (a hemizygous deletion) or both copies (a homozygous, or full, deletion) of

genetic material. A homozygous deletion is likely to be more deleterious than a hemizygous deletion, because both copies of a given genomic sequence, which may include genes, are deleted. Duplication CNVs exist as three copies, four copies or more than four copies. Table 1.1 illustrates current CNV nomenclature. Note that the X chromosome follows the same convention in females, but in males a deletion or duplication event result in zero or more than one copy of genomic material respectively.

Total DNA copy number	**Description**	**Associated genotypes (each genotype is arbitrarily denoted A and B)**
0	Deletion of two copies (homozygous deletion)	Null
1	Deletion of one copy (hemizygous deletion)	A, B
2	Normal state	AA, AB, BB
2	Normal state with loss of heterozygozity (AB)	AA, BB
3	Single copy duplication (hemizygous duplication)	AAA, AAB, ABB, BBB
4 (+)	Double copy duplication (homozygous duplication)	AAAA, AAAB, AABB, ABBB, BBBB

Table 1.1. Copy number variation nomenclature for autosomal chromosomes.

CNVs occur with differing frequencies throughout populations, and are divided for convenience into those defined as common (occurring in more than 1 in 100 individuals in a population) and rare (less than 1 in 100)(Pinto et al., 2010). When studying associations with disease states, the rarest CNVs are logically of the most interest, since they are more likely to be evolutionarily under negative selection pressure (and thus conferring disadvantage, possibly disease).

De novo CNVs are defined as CNVs that are seen in a child but not in their biological parents and are formed in the cells that give rise to sperm and eggs, or shortly after conception has taken place. CNVs may also be referred to as 'singleton' CNVs, which refers to a CNV that is only seen once in a dataset. Such CNVs are also, presumably, rare in the general population.

A CNV, similar to any genetic variant, may be described in terms of 'penetrance', which is defined as the proportion of individuals carrying a particular variant that also express an associated phenotype(Lobo, 2008).

Much of the focus of current research is on rare CNVs and their aetiology in neuropsychiatric disorders. Whilst common CNVs may also play a role, their association is complicated by the problem of different frequencies in different ethnic groups, and current evidence suggests that their effects are, at best, modest(Conrad et al., 2010).

1.6 Copy Number Variants in Psychiatric Disorders

1.6.1 Learning Disability

Learning disability, which is also termed intellectual disability or mental retardation, covers a range of disorders that present with life-long cognitive deficits that become apparent in the years after birth. The range of cognitive deficits seen within individuals with learning disability is wide. Some individuals demonstrate broad, even cognitive deficits across domains, whilst others have circumscribed deficits with other intellectual functions being retained. Some patterns of intellectual and physical disability appear repeatedly and have been

associated with a particular genetic abnormality, with the most obvious example being Down syndrome. In many instances these syndromes are caused by copy number variation in the affected area, and we turn to two such examples first. For each syndrome described, we also cite the reference number for the associated disorder in the comprehensive web resource for genomic disorder Online Mendelian disorders in Man (OMIM).

Smith-Magenis syndrome (OMIM 182290) and Potocki-Lupski syndrome (OMIM 610883) represent good examples of disorders caused (in most cases) by copy number variants; a deletion and a duplication respectively. In these two disorders the same region of the genome (17p11.2) is implicated, and in this sense both the CNV and associated disorders are termed 'reciprocal'.

90% of patients with Smith-Magenis syndrome exhibit a sporadically occurring CNV deletion involving the gene RAI1 (retinoic-acid induced 1). The remaining 10% have point mutations in this gene(Elsea & Girirajan, 2008). They are particularly liable to severe self-injurious behaviours such as head banging, skin picking and wrist biting, beginning early in life. Some individuals have normal IQ and may manage sufficiently well at school to avoid detection. This striking range of severity suggests that other genetic factors probably play a part in determining the overall outcome of the disorder, although the RAI1 gene is a primary cause.

Potocki-Lupski syndrome, as the reciprocal disorder of Smith-Magenis syndrome, is caused in most cases by a duplication CNV in the same region of chromosome 17p11.2. Potocki-Lupski syndrome often presents with a milder phenotype than

Smith-Magenis syndrome, with infants demonstrating failure to thrive, sleep apnea, cardiovascular abnormalities, mental retardation and, interestingly, autistic features(Potocki et al., 2007).

Williams-Beuren syndrome (OMIM 194050) occurs in approximately 1 in 7,500 to 1 in 20,000 births and is characterised by mild to moderate learning disability, heart and vascular abnormalities and facial dysmorphism. The disorder is strongly linked to a 1.5-1.7MB CNV deletion of 7q11.23(PEOPLES et al., 2000). Numerous genes are present within this region, and the LIMK1 gene, which is highly expressed in brain, may be implicated in the unique cognitive profile of this disorder. Patients with Williams-Beuren syndrome demonstrate intact language and facial processing skills but profound visuospatial processing deficits. They have a distinct, hypersociable personality and often have notable musical talents and fearfulness to certain sounds(Martens, Wilson, & Reutens, 2008).

The reciprocal duplication syndrome of Williams-Beuren syndrome is also seen (although it lacks an eponym) and fascinatingly this syndrome is most robustly associated with marked language delay and autistic features (from a certain point of view, the reciprocal personality of that seen in Williams-Beuren syndrome), along with a range of physical health problems(J. S. Berg et al., 2007; Van der Aa et al., 2009). Again, the range of severity is wide.

The observation of a reciprocal disorder caused by variants in the same region of the genome is not uncommon, and further examples are given in table 1.2.

CNV Locus	Del/Dup	Eponymous syndrome	OMIM Number
17p11.2	del	Smith-Magenis syndrome	182290
	dup	Potocki-Lupski syndrome	610883
7q11.23	del	Williams-Beuren syndrome	194050
	dup	None	609757
22q11.2	del	Velocardiofacial/DiGeorge/ Shprintzen/CATCH 22 syndromes	188400
	dup	None	608363
16p11.2	del	None	611913
	dup	None	611913

Table 1.2. Examples of four reciprocal syndromal CNVs in the genome.

Reciprocal deletions and duplications of certain parts of the genome occur due to a process called non-allelic homologous recombination(Inoue & Lupski, 2003). During this process, which usually occurs during the formation of gametes, reciprocal copy number events are created in exactly equal numbers. The observation of a somewhat milder phenotype in duplication syndromes is usual, and may indicate that duplication events are in general, and perhaps logically, less deleterious then deletion events. Duplicated genetic elements that subsequently diversify from each other are a known evolutionary mechanism(J. Zhang, 2003).

1.6.2 22q11.2 Deletion Syndrome

A wide phenotype, from severe mental retardation to a picture which escapes clinical detection is also seen in the 22q11.2 deletion syndrome, which deserves special mention as it is probably the most common genetic deletion syndrome, and frequently manifests with psychiatric disorder(Robin & Shprintzen, 2005).

The syndrome is caused by a 1.5-3MB deletion CNV of chromosome 22q11.2. It is seen in approximately 1 in 2,000 to 1 in 9,000 individuals(Goodship, Cross, LiLing, & Wren, 1998; Montcel & Mendizabai, 1996; Shprintzen, 2001). As with many disorders with a variable phenotype severity, the frequency of the CNV may be under-ascertained in the population. The clinical manifestation of 22q11.2 deletion syndrome is highly variable, with velocardiofacial syndrome (VCFS), CATCH 22 syndrome and DiGeorge syndrome being a few examples of syndromes described in individuals with the deletion(Shprintzen et al., 1978). As well as learning disability, cardiac abnormalities, facial and palatal deformities are frequently seen(Robin & Shprintzen, 2005). This deletion syndrome is particularly significant because it has been known for over a decade that it is associated with a notably high incidence of schizophrenia(Murphy, Jones, & Owen, 1999). ADHD, autism and bipolar disorder have also been associated with the deletion(Karayiorgou, Simon, & Gogos, 2010). Some individuals demonstrate no discernable abnormality.

The region covered by the deletion encompasses approximately 60 genes, depending on the size of the variant, and most are expressed in brain(Maynard et al., 2003). This region of the genome is highly prone to rearrangement, perhaps explaining the high frequency of this mutation(Saitta et al., 2004).

A reciprocal duplication syndrome of 22q11.2 is also seen, and logically this should be as prevalent as the deletion syndrome. However it is not often observed in clinical populations, and relatively few studies of it exist(La Rochebrochard et al., 2006; Ou, Berg, Yonath, & Enciso, 2008; Portnoī, 2009;

Yobb et al., 2005). The phenotype appears to be much milder, with many individuals being largely normal. This milder, highly variable phenotype probably explains its low detection rate.

1.6.3 Autism

Autism is a complex neuropsychiatric disorder comprising a core phenotype of delayed language development, ritualised and restricted behavioural patterns and deficits in reciprocal social interaction(American Psychiatric Association, 1994; World Health Organisation, 1992). Evidence from twin studies suggests that is has a high heritability of over 80%(A. Bailey et al., 1995), and 5% of cases are associated with known genetic syndromes such as Rett syndrome, Fragile X syndrome and tuberous sclerosis(Freitag, 2007). It typically presents before the age of 3 although less severe forms may become apparent later in life and the disorder is usually conceptualised as a spectrum, with milder cases, often defined as Asperger's syndrome or atypical autism, presenting with unimpaired intellectual function and speech development but with prominent problems in social interaction and ritualised or restricted ranges of behaviours and interests remaining relatively prominent(A. Bailey, Phillips, & Rutter, 1996). Males are 4 times more likely to be diagnosed with autism than females.

5% of autistic patients have a chromosomal abnormality visible under the light microscope(Vorstman et al., 2005). Added to this, linkage analysis studies have found evidence of association of over 20 chromosomal regions with autism, giving an early indication of its complex genetic aetiology(Klauck, 2006).

Sebat et al.(Sebat et al., 2007) were one of the first groups to publish findings from their use of a DNA microarray looking for deletions and duplications of genomic material in autistic individuals. 165 families (118 with only one affected child and 77 with multiple cases), and 99 control families were collected. They found 15 CNVs in 14 of 195 autistic cases, mostly comprised of deletions (12 of 15) whereas in 196 controls they found only 2 CNVs in 2 individuals, (both duplications). Statistically this was a highly significant difference. 10% of the children in families with only 1 affected child (12 out of 118) demonstrated a CNV that could not be found in either parents (a de novo mutation). In contrast, only 3% (2 out of 77) of families with cases of autism spanning generations had a de novo CNV, and 1% (2 out of 196) of control families. Amongst the CNVs they identified, two overlapped with genomic regions already implicated in autism (15q11-13 and 16p11.2).

They made the interesting observation that de novo CNVs were enriched in females in their cases. Since autism is diagnosed less frequently in females than males (and therefore being female can be construed as protective), they postulated that the CNVs they observed in this group were more likely to be contributing to the development of autism. They also postulated that de-novo CNVs may be more associated with sporadic cases of autism than with cases where multiple family members are affected.

Weiss et al.(Weiss et al., 2008) have also genotyped a large number of families with autism. Similarly to Sebat et al., they found an enrichment of the deletion CNV at 16p11.2 in their sample. They also observed the reciprocal duplication in

this area in three families within their case sample, with one family demonstrating 4 affected male siblings with autism who all had the duplication. However in two families this duplication was also present in the unaffected parents (making its aetiological significance less compelling).

In a separate cohort they also found the 16p11.2 duplication in 3 females with bipolar disorder. They did not find this duplication in their control cohort. This provides evidence for the involvement of the 16p11.2 region in autism, but the deletion appears to be more significant than the duplication and the observation of this CNV in a bipolar dataset suggests CNVs in this region are unlikely to segregate solely with autism, indeed they may be seen in many other disorders, as well as normal controls, too.

In addition 5 large duplication CNVs in the 15q11-13 region of the genome were observed (classically this region is associated with Prader-Willi and Angelman syndromes). Of these five events, three were de novo.

6 families were observed to carry a deletion covering the neurexin-1 gene. Neurexin-1 is one member of the neurexin family of proteins which, because of their role in supporting synaptic function, are of particular interest in autism. However of the 6 families where the deletion was seen, only 2 members were affected with autism.

Other CNVs in different regions of the genome previously associated with autism were also observed in this study, but the numbers observed in this sample were not sufficient for a statistically significant association to be shown.

Weiss et al. also commented in this study on the large number of CNVs that appeared to be benign variants. Disentangling the CNVs which contribute to a disorder from those that are incidental occurrences is challenging. Simple single CNV-single disorder associations do not take account of interaction effects between CNVs (or other genetic variants). Some CNVs may be irrelevant to the disorder being studied (although not to another). It is even possible (but probably less likely) that some CNVs are protective. Current studies in CNVs are not sufficiently sophisticated or powered to analyse for these complex effects.

Marshall et al.(Marshall et al., 2008) studied the occurrence of rare and de novo CNVs in 427 families with cases of autistic spectrum disorders compared to 500 controls. Using past evidence, they focussed on CNV regions previously implicated in autism. They found that 27 cases had de novo CNVs. 4 cases had identical CNVs at 2 loci and 27 unrelated cases had similar CNVs. They found CNVs in the regions 22q11.2 and 15q11-13 and CNVs in genes that have previously been associated with autism, for example *NLGN4* (neuroligin 4), *SHANK3* (SH3 and multiple ankyrin repeat domains 3) and *NRXN1* (neurexin 1). They detected two cases of the 22q11.2 deletion CNV (both de novo) in two brothers affected with autism. Interestingly these two brothers were each at opposite ends of the autistic spectrum. One was much more impaired than the other. Duplications at 22q11.2 were also found, but, in keeping with the idea that a duplication in this area is less deleterious, in one family the duplication was found in the normal father, but not in his autistic son. CNVs at 16p11.2 were also detected in autistic cases, but also in normal controls.

The group went on to look for these CNVs in a further set of 1,152 matched controls. They did not find CNVs found in autism cases in any of the controls. Overall from this data the authors estimated that approximately 7% of cases with an unknown aetiology harbour de novo CNVs detectable by microarrays. This is a significant number and suggests that the routine use of microarrays in clinical scenarios where autism is suspected may be appropriate.

1.6.4 Schizophrenia

Schizophrenia is a common (~0.8% lifetime incidence) psychiatric disorder with a substantial heritable component estimated at around 80% and with first degree relatives of sufferers having a 10-fold increased risk of also developing the disorder (for a review see(Gottesman, 1991)). Whilst it can occur at any age, even in childhood, it is most commonly seen in early adulthood and is characterised by problems with perception, inferential thinking, goal-directed behaviour and emotional expression(Andreasen, 1995) as well as, cardinally, a lack of insight into the disorder. No one symptom or sign is common to all individuals diagnosed with the disorder. Many follow a chronic course of illness suffering substantial psychosocial deficits and increased morbidity and mortality for the rest of their lives, especially from suicide(Harris & Barraclough, 1998). Whilst past genetic studies involving linkage, single nucleotide polymorphisms and candidate genes have discovered some interesting regions in the genome, often these have not been replicated (for a review see, for example(Schwab & Wildenauer, 2009) and (Sullivan, 2005)). A substantial proportion of heritability remains to be explained.

The first CNV studies in schizophrenia were published in late 2007 and early 2008. The first report by Kirov et al.(Kirov et al., 2007) used CGH arrays to investigate 93 individuals with schizophrenia compared to 372 controls. They filtered out all CNVs found in the control group or found in the Database of Genomic Variants (http://projects.tcag.ca/variation/) to leave 13 CNVs for further analysis. They identified two CNVs which were very likely to be aetiologically significant. The first was a 1.4MB duplication in chromosome 15q13.1, covering the gene *APBA2* (Amyloid beta A4 precursor protein-binding family A member 2), *NDNL2* (necdin-like 2) and *TJP1* (Tight junction protein ZO-1) genes and which was found in neither biological parent (a de novo event). The second was a 250kb deletion of 2p16.3 covering the gene *NRXN1* (neurexin 1), although this was also found in the asymptomatic mother.

Walsh et al.(Walsh et al., 2008) studied 150 patients with schizophrenia and 268 ethnically matched controls, genotyped on CGH arrays and followed up with SNP microarrays, with a replication sample consisting of 92 patients with childhood-onset schizophrenia. They found a three-fold enrichment of CNVs in gene-coding regions in the schizophrenia cohort and those with an onset of illness less than 18 years were more than 4 times as likely to carry a CNV as controls. 28% of the childhood onset schizophrenia cohort (which is conceptualised to represent a particularly severe form of the disease) carried a CNV over a gene-coding region, again significantly greater than the control sample. Of the CNVs that were detected, the authors found that they were significantly more likely than chance to cover genes known to be associated with central nervous system development and function. Neuregulin signalling, ERK/MAPK signalling, synaptic long-term

potentiation, axonal guidance signalling, integrin signalling, and glutamate receptor signalling were particularly associated. Contrastingly, in the control group the CNVs observed covered genes involving any biological pathway. The authors looked at a number of gene-coding CNVs in the case cohort in more detail, for example ERBB4, and showed that the resulting transcribed protein was aberrant and unlikely to be functional.

Xu et al.(Xu et al., 2008) published an analysis of 152 individuals with schizophrenia and their biological parents compared against 159 controls drawn from the same population. The advantage of this approach is that, by looking at the biological parents of the cases, the group could determine whether the CNVs observed in the cases were de novo.

They found 15 cases with CNVs where the CNV did not occur in the parent (de novo) as opposed to 2 in controls, a significant enrichment of approximately 8 fold. 2 cases carried 2 de novo CNVs each. To assess whether or not de novo CNVs might be important in the aetiology of schizophrenia they then went on to look for de novo CNVs in a further cohort of 48 cases of schizophrenia who also had a first or second degree relative who had the disorder. In contrast, they did not find de novo CNVs in this cohort. They also found that sporadic cases of schizophrenia were only 1.5 times more likely than unaffected controls to harbour an inherited rare CNV.

Taken together this suggests that de novo CNVs are more likely to explain sporadic cases of schizophrenia than cases where there is a family history of the disease, a similar picture to autism. It also suggests that only a small proportion

of rare inherited CNVs might be expected to contribute to the pathogenesis of sporadic cases. They did not find any association of rare CNV events with age of onset, presence or history of learning disability or parental age at birth. 3 cases of the 22q11.2 deletion were identified in this cohort, of which two had mild facial dysmorphism.

In September 2008 the international schizophrenia consortium reported results of their analysis in 3,391 patients with schizophrenia and 3,181 ethnically matched controls in the journal Nature(International Schizophrenia Consortium, 2008). They found that the overall burden of CNVs was increased 1.15 fold in their case cohort in comparison to the control cohort. They also found that large CNVs (over 500kb), CNVs that occurred rarely in the overall sample, and CNVs that covered genes, were particularly enriched in the cases of schizophrenia.

The general increase in burden of rare CNVs reflected in part the observation of large, rare deletion CNVs in a minority of cases but no controls. Whilst statistical tests of association of these large CNVs mostly failed to reach significance levels, probably because of a lack of statistical power, similar CNVs to those found in other disorders such as autism were observed, which taken together with other studies is an interesting finding.

The 22q11.2 deletion was found to be significantly associated, but the group also reported new associations with deletion CNVs at 15q13.3 and 1q21.1. The 15q13.3 deletion had also been associated with mental retardation and seizures in a paper earlier that year(Sharp et al., 2008) and the region at 1q21.1 has

previously been associated with schizophrenia by linkage(Brzustowicz, Hodgkinson, Chow, Honer, & Bassett, 2000).

In a further large study, Stefansson et al.(Stefansson et al., 2008) initially looked for de novo CNVs in an exploratory population sample, and then tested these 66 regions for association in a sample of 1,433 cases of schizophrenia against 33,250 controls. Those regions found to have suggestive association were followed up in a replication cohort comprised of 3,285 cases and 7,951 controls. CNV deletions at 1q21.1, 15q11.2 and 15q13.3 were identified as candidates for association in the first sample and all three of these deletions showed significant association in the follow up sample.

The 22q11.2 deletion was present in 8 out of 3,838 cases but notably absent in all controls. This lends support to the idea that schizophrenia is polygenic, but that some genetic variants confer more risk than others. Interestingly in this study, the authors also had access to clinical data, so they were able to study whether or not various clinical indicators were different in those individuals with CNVs. They found no differences in clinical response to treatment and there was no sex bias or tendency for those with a CNV to have a family history of schizophrenia. The authors noted that three cases carrying the 1q21.1 deletion had learning disabilities and of the eight controls that had the deletion, two had dyslexia. The 'multiple hit' hypothesis is hinted at here. No one individual variant is either necessary or sufficient to cause schizophrenia (although they may cause mild sub-clinical problems on their own), but together, multiple variants may act to predispose an individual strongly towards developing the disorder.

McCarthy et al.(Shane E McCarthy et al., 2009) investigated the role of recurrent CNVs at 16p11.2 in schizophrenia, working on the basis of past findings implicating this region in autism(Marshall et al., 2008; Weiss et al., 2008). This research group had also previously detected this variant in two cases of childhood-onset schizophrenia(Walsh et al., 2008). Two cohorts of schizophrenia cases were analysed, matched against suitable control sets. In the first cohort they found the 16p11.2 duplication CNV in 12 out of 1,906 cases (0.63%) but only 1 out of 3,971 controls (0.03%). In the second cohort they found this duplication in 9 out of 2,645 (0.34%) cases but only 1 out of 2,420 controls (0.04%). This represents a 14.5 fold increased risk of schizophrenia for those carrying the CNV.

Interestingly, the deletion CNV at 16p11.2 was not associated with schizophrenia, however in a meta-analysis of previous reports the authors showed that the deletion was in fact more associated with developmental disorders and autism rather than schizophrenia, whilst the duplication might be associated with all of autism, schizophrenia and bipolar disorder.

At first glance the results of the analysis of this region of 16p seem confusing. Deletions and duplications at 16p11.2 appear to be associated with autism, whilst only the duplication is associated with schizophrenia and bipolar disorder. Some individuals with 16p11.2 CNVs are apparently normal. Deletion CNVs, because they reduce gene dosage by 50%, are more likely to affect overall function than duplication CNVs, which increase gene dosage by 33.3%. Interestingly, duplication events appear to be critical in human evolution, where

divergence of a duplicated gene from its original allows the formation of genes and proteins with different functions (J. Zhang, 2003). If a gene copy is deleted this process cannot happen. This observation sheds some oblique light on why a more severe disorder such as autism may be only associated with the deletion CNV. Schizophrenia and bipolar disorder, being of later onset and also, usually, being treatable, can be considered less severe and thus perhaps the association with the duplication is more likely. But this does not explain why 1) normal controls are seen with the CNV and 2) why the duplication CNV is also seen in autism. For a complex disease such as autism, it is more likely from current evidence that multiple deleterious variants are required, each contributing a small, and probably clinically undetectable liability. Thus normal individuals may harbour the CNV, but because they do not possess other relevant variants, go clinically undetected. Those with an autism diagnosis more probably manifest a range of unfavourable variants, perhaps with interacting effects with the 16p11.2 CNV. With this in mind, the original result is perhaps more explicable, and at the same time we receive a glimpse of the genetic complexity of these disorders.

Need et al.(Need et al., 2009) studied CNVs in 1,013 cases of schizophrenia compared to 1,084 controls. They initially looked at CNVs larger than 2MB in size and found 8 deletions, all in cases, and 9 duplications, of which 6 occurred in cases and 3 in controls. Large deletions were statistically more likely to occur in cases than controls, but not duplications. They then went on to look for deletion events larger than 2MB in a further cohort of 1,547 controls screened with an extensive battery of cognitive tests. They did not find any deletions larger than

2MB in this group, further underlining the association with the schizophrenia cohort. They observed 4 of the large deletions in the 22q11.2 region, 3 in the 1q21.1 region and 2 new CNVs, one in 16p13.11 (and within which two further cases had a smaller CNV) and further large deletion on 8p22. They also saw duplications larger than 2MB in 15q11.2-13.3, extending across the Prader-Willi/Angelman syndrome critical region and the gene APBA2, which had previously been reported in another case of schizophrenia by Kirov et al.(Kirov et al., 2007). They commented that although duplications were also seen in controls, very large duplications (>3MB) were not, and segregated entirely with cases, suggesting that although duplications may not be as deleterious as deletions, very large duplication events may be significant in schizophrenia. The also investigated common CNVs, which were arbitrarily defined, using an approach capitalising on the fact that some SNPs tag CNVs (that is, a particular SNP allele segregates with the CNV). They did not find any significant associations in their data.

Levinson et al.(Levinson et al., 2011) studied 3,945 subjects with schizophrenia or schizoaffective disorder compared to 3,611 screened controls and compared the rates of rare CNVs between groups. The use of screened controls in this study adds power to the groups ability to detect an association, as it is reasonable to expect controls screened for an absence of medical problems to harbour less CNVs than cases. The group confirmed previous associations found for deletions at 15q13.3, 1q21.1 and 22q11.2 as well as duplications at 16p11.2. They identified a new association region at 3q29. Seven cases had a deletion in this region, but no controls. This CNV had previously been found in 14 individuals in

a sample of 14,698 individuals with mild-moderate mental retardation, and a variety of other phenotypes, including autism(Ballif et al., 2008). Overall this paper provides strong independent support for the role of certain rare CNVs in schizophrenia.

Buizer-Voscamp et al.(Buizer-Voskamp et al., 2011) investigated CNVs over gene coding regions in 834 Dutch schizophrenia patients and 672 Dutch control subjects screened for an absence of psychiatric symptoms and history. They found that their cases had a similar burden of CNVs (2.1 per sample) to controls, however when they stratified their CNVs by type (deletion or duplication) and size (50-500kb, 500kb-1MB and ≥1MB) they found significantly more deletion CNVs in cases than controls across size ranges. They also found CNVs overlapping with those previously found in other studies (1q42 and 22q11.2) and a new locus at 5q35.1.

The CNVs discussed so far are large, usually spanning hundreds of thousands or millions of base pairs. The advantage of studying such CNVs is that they are easy to detect with current array technology, and stand out as candidates for disease association. The disadvantage of studying such large variants is that they usually encompass large numbers of genes. One or more than one gene may be relevant to a particular disorder, but without more focussed study it is difficult to tell which gene may be playing an aetiological role. In this case, smaller CNVs that cover or interrupt individual genes are more informative, although they can be harder to detect.

The gene *NRXN1* (neurexin 1) encodes a presynaptic neuronal cell surface protein that mediates neurotransmission by interacting with neuroligins. Disruptions of this gene impair normal synaptic function (Sudhof, 2008). CNVs in this gene were shown by Kirov et al. in two siblings with schizophrenia and an analysis of this gene within a sub-cohort of schizophrenia patients in 2008 found significant associations between CNVs affecting exons of this gene and schizophrenia (Kirov et al., 2007; Rujescu et al., 2009). Walsh et al. also found a deletion of neurexin 1 affecting identical twins with early onset schizophrenia (Walsh et al., 2008). Marshall et al. and Weiss et al, amongst others, have also linked variants in neurexin 1 to autism (Feng et al., 2006; Kim et al., 2008; Marshall et al., 2008; Morrow et al., 2008; Szatmari et al., 2007; Weiss et al., 2008) and a further study has linked this gene to mental retardation (Friedman et al., 2006). The neurexin-1 gene is large, and the CNVs demonstrated in the gene (usually deletions) can variously affect different parts of the gene. CNVs that disrupt exons (protein coding parts of genes) appear to be most associated with disease, as might be expected, but other CNVs may also be relevant.

A complicated picture for even a single gene emerges here. Neurexin 1 encodes alpha and beta subunit proteins formed from transcription via different promoters, and multiple isoforms of each alpha and beta variant are produced by splicing of the messenger RNA. A disrupting CNV within neurexin 1 is likely to disrupt the balance of splice variants produced at different sites in the CNS, and probably subtly affects synaptic neurotransmission, with the extent of disruption being dependent on the precise nature of the CNV involved. In a review of this gene and meta-analysis of association studies Kirov at al. showed the CNV to be

present in 17 out of 8,798 cases (0.19%) but only 17 out of 42,054 controls (0.04%) ($p=1.3x10^{-5}$) with deletions of more than 100kb even more strongly associated with disease ($p=3.7x10^{-6}$)(Kirov, Rujescu, Ingason, Collier, O'Donovan, & Owen, 2009b). So, whilst we have quite compelling evidence of association, neurexin-1 CNVs only account for a tiny proportion of overall liability to disorder, and most cases will not exhibit a CNV in this gene. Nonetheless here is an important small piece to a much larger puzzle.

1.6.5 ADHD

Attention deficit hyperactivity disorder (ADHD) is a common childhood psychiatric disorder that in a substantial proportion of cases continues into adulthood(Biederman, 1998). It is more often seen in boys and comprises extreme degrees of impulsivity, inattentiveness and hyperactivity. Whilst it remains a somewhat controversial diagnosis in some circles, a recent meta-analysis estimated the worldwide pooled prevalence at 5.29%(Polanczyk, de Lima, Horta, Biederman, & Rohde, 2007). The term 'hyperkinetic disorder' is often used outside of North America, but the diagnostic criteria are similar(World Health Organisation, 1992). ADHD has substantial heritability estimates from twin and adoption studies of around 76%(Franke, Neale, & Faraone, 2009). However the genetic aetiology remains obscure.

Elia et al.(Elia et al., 2010) genotyped 335 cases of ADHD and their parents and 2,026 healthy controls to study CNVs. They detected 222 inherited CNVs in cases that were not present in any of the controls. 28 covered 22 genes previously implicated in other disorders such as autism and schizophrenia. 11 CNVs had

previously also been associated with autism and schizophrenia. Whilst the authors, in a pathway analysis, found an over-representation of genes involved in learning, central nervous system development, hindbrain development and cell adhesion, they found no difference in mean CNV count and CNV size between cases and controls. This suggests that cases of ADHD, whilst not having a higher CNV burden in general, do have a focussed set of CNVs occurring over genes occurring in particular biological pathways.

In a further large study of large, rare CNVs in 410 children with ADHD compared to 1,156 controls, Williams et al.(N. M. Williams et al., 2010) detected a highly significant difference of 57 CNVs in their case set, equivalent to 15.6% of cases, and 78 in the control set, equivalent to 7.5%. Of particular note in this study was that some children had intellectual disability (IQ < 70). However even when cases with intellectual disability were excluded, 12.5% of the cases still harboured a CNV, maintaining statistical significance. The average IQ of children in the case group was 89, perhaps suggesting that more CNVs than this might actually be attributable to general intellectual ability. However, the overlap between different neurocognitive problems (such as ADHD) and general intellectual ability (measured by IQ) is particularly hard to measure, and may in any event not be a meaningful distinction. This reinforces the notion that CNVs can confer a broadly increased risk of neuropsychiatric illness (that may, in part, be mediated by general intellectual ability) rather than contributing to any particular disorder on its own.

Lesch et al.(Lesch et al., 2011), in a smaller and more focussed study, considered 99 children and adolescents with severe ADHD. Two thirds of these children came from families with at least two members affected with ADHD, 8 patients were from extended families with high numbers of sufferers of ADHD and the remainder (24 cases) were sporadic. In contrast to Williams et al. they excluded any case with an IQ < 80 in this study, and any case with a comorbid diagnosis of autism or schizophrenia. They found 17 CNVs in their cohort; 4 deletions and 13 duplications. 2 CNVs were de novo. Many of the CNVs contained genes that may be associated with ADHD aetiology, and the study provides a good basis for more focussed research on the neurobiological changes that these CNVs may elicit.

1.6.6 Bipolar Affective Disorder

Bipolar affective disorder occurs in about 1% of the general population, affecting males and females equally, and is characterised by episodic periods of mania and depression (or less commonly mania alone) usually accompanied by disturbances in behaviour and thinking (World Health Organisation, 1992). In contrast to schizophrenia, where residual deficits are common, recovery is usually complete between episodes (Judd & Schettler, 2010). Most twin and adoption studies indicate a heritability of about 80% and the lifetime risk for the disorder in monozygotic co-twins of a sufferer of bipolar disorder is about 40-70%, and first degree relatives about 5-10%(Craddock & Jones, 1999). SNP GWAS of bipolar disorder have met with some success, with the Psychiatric GWAS Consortium Bipolar Disorder Working Group, in a meta-analysis 11,974 cases of bipolar disorder compared to 51,792 controls, confirming an association with the alpha subunit of the L-type voltage gated calcium channel (CACNA1C)

and a new SNP in an intron of a human homolog of the Drosophila pair-rule gene ten-m (ODZ4)(Sklar et al., 2011). Previous studies had also identified SNPs in the neurocan (NCAN) gene(Cichon et al., 2011) and the ankyrin 3 (ANK3) gene(Schulze et al., 2009).

Zhang et al.(D. Zhang et al., 2009a) compared the frequency of rare CNVs detected in 1,001 bipolar affective disorder cases against 1,034 controls. Looking specifically at singleton deletions (deletion CNVs occurring only once in the entire sample) they found that 16.2% of cases contained a singleton deletion CNV compared to 12.3% in controls. Some singleton CNVs covered genes already implicated in schizophrenia, specifically GRM7 and LARGE. Bipolar disorder and schizophrenia are considered to be distinct yet overlapping disorders, and so this is an interesting and plausible finding. GRM7 encodes a metabotropic glutamate receptor expressed in brain and LARGE encodes a gene which has previously been associated with mental retardation (MDC1D)(Longman, 2003).

Singleton CNVs are an interesting group of variants, and a logical group to focus on. They are defined as occurring only once in a dataset, and they thus are more likely to represent variants that occur rarely in a population. Rarely occurring variants may be more likely to be functional.

Grozeva et al.(Grozeva et al., 2010) have also studied CNVs in a sample of 1,697 bipolar cases compared with 2,806 unscreened controls. In contrast to Zhang et al., they found significantly *less* deletion CNVs in the cases of bipolar disorder than controls. Of note however is that the control group in this study is unscreened, whereas Zhang et al. screened their control group for an absence of

neuropsychiatric disorder. The two studies are difficult to compare in this context.

McQuillin et al.(McQuillin et al., 2011) focussed on 546 cases of bipolar disorder compared to 517 controls screened for lifetime absence of psychiatric disorder. Similar to Grozeva et al, but in contrast to Zhang et al, they found no evidence to suggest an increased burden of rare CNVs, including singleton CNVs. They also found that, overall, bipolar cases tended to have a lower burden of CNVs compared to, in this case, screened controls. Notwithstanding this they found instances of CNVs previously seen in other CNV association studies. 5 cases had a CNV on chromosome 19 covering a variety of zinc finger domain containing proteins that had previously been identified by the International Schizophrenia Consortium's paper(International Schizophrenia Consortium, 2008). 3 cases had a CNV on chromosome 1 identified by Zhang et al.(D. Zhang et al., 2009a). 4 cases had CNVs on chromosome 10 that had also been seen in Grozeva et al.'s paper(Grozeva et al., 2010).

Priebe et al.(Priebe et al., 2011) studied 882 cases of bipolar disorder compared to 872 population controls. They found that the frequency of samples with rare duplications and the average total length of singleton deletions was significantly increased in cases with an age of onset of disorder of ≤21 years when compared to controls. They also found evidence for a significant over-representation in cases of two common CNV regions in 10q11 and 6q27. Comparing the frequency of CNVs in regions previously implicated in psychiatric disorders, they did not find a significant increase when cases were compared to controls in any region

however, like many studies, the rarity of the CNVs themselves suggests a lack of power to detect an association.

Malhotra et al.(Malhotra et al., 2011) have recently published a study looking at a cohort of 788 trios, including 185 with bipolar diagnoses and 177 with schizophrenia diagnoses. They found that the frequency of de novo CNVs was significantly higher in the bipolar diagnosis cohort, with an odds ratio of 4.8, and a particular enrichment in those with a younger age of onset (<18). Within the schizophrenia cases a similar picture for de novo CNVs was also observed.

1.6.7 Major Depressive Disorder

During the course of this thesis, two studies of CNVs in major depressive disorder have been published, in addition to our own(Rucker et al., 2011).

Glessner et al.(Glessner et al., 2010) analysed 1,693 cases of depression and 4,506 controls genotyped on the Perlegen 600k platform. They found significant association between cases and a duplication at the SLIT1 gene on chromosome 5. This study was unfortunately hampered by the microarray platform which, because the polymerase chain reaction in the assay is run to saturation, tends to have high signal to noise ratios that make calling CNVs accurately difficult.

Degenhardt et al.(Degenhardt et al., 2012) genotyped 604 cases of major depressive disorder and 1,643 controls on Illumina SNP arrays. They found evidence that CNVs in 4 regions of the genome, 7p21.3, 15q26.3, 16p11.2, and 18p11.32, were associated with their depressed cases. Reciprocal CNVs at

16p11.2 have already been discussed in this section, associated with many different psychiatric disorders.

These studies will be discussed in more detail in our final chapter.

1.6.8 Common CNVs

Common CNVs, which are usually called copy number polymorphisms (CNPs) are defined variably by different groups, but usually as occurring in more than 1%, 5% or 10% of the population (Conrad et al., 2010; Craddock et al., 2010; Grozeva et al., 2010; McQuillin et al., 2011; Rucker et al., 2011). Depending on the definition used they account for a varying but significant proportion of variation within the genome, but have not been as intensively studied as rare and de novo CNVs, probably because they are intuitively less attractive candidates for explaining disease susceptibility. Some groups have argued that CNPs may in fact account for most of the susceptibility to complex diseases(Risch & Merikangas, 1996). Others propose that both CNPs of small effect and rarer CNVs of larger effect play interacting roles(Rucker & McGuffin, n.d.; Uher, 2009).

Previous work in genome wide association studies using single nucleotide polymorphisms (SNPs) observed that most CNPs are 'tagged' by single nucleotide polymorphisms(Craddock et al., 2010). That is, the presence of a CNP can be inferred if a particular allele at a particular SNP is seen. Using this observation, Conrad and Pinto were able to analyse the origins and impact of CNPs and show that, at least with current methods for analysis and detection, that they are unlikely to contribute to a substantial proportion of heritability in neuropsychiatric disorders(Conrad et al., 2010). This is not a surprising result, as

no particular form of genetic variation has yet been shown to substantially contribute to complex disease susceptibility. However it does place CNPs in context, and inform us that, as with CNVs, SNPs and other forms of genetic variation, our expectations should be modest.

1.7 Conclusion

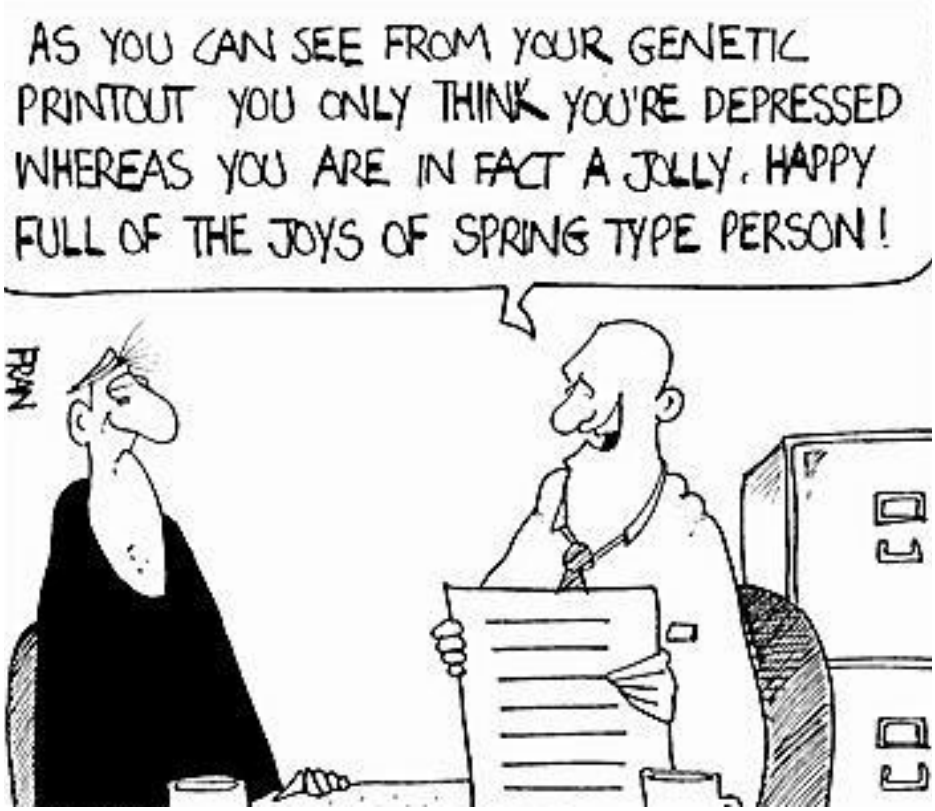
Multiple lines of evidence exist that CNVs, especially those that are rare and de novo, contribute to susceptibility for complex neuropsychiatric diseases like autism, learning disability and schizophrenia. Most datasets in existence have been collected based on the aggregation of groups of individuals who share similar diagnoses or characteristics. In this context the aetiological association of CNV with disease state is natural, but is likely to lead to an over-simplification of a more complex picture than a straight forward disease-to-variant association. Nature has no need to respect our own phenotypic delineations, and so this appears to be the case at this level of genetic variation. It may be more reasonable to conceive from the evidence so far that rare and de novo copy number variants are associated with a general liability to neuropsychiatric, and other, outcomes, rather than specific diseases. Large, rare deletion CNVs may be associated with more severe disorders than duplications, but even here the spectrum of phenotypic variation is surprisingly wide, suggesting an interacting concert of protective and deleterious variants in addition to those observed.

De novo events may partly explain some cases of sporadic disease occurring in families with no prior history, whilst transmissible events may also contribute to

disease. The observation of similar variants in the normal parents of affected probands does reinforce the concept of variable penetrance, and probably lends evidence to a multiple-hit hypothesis of aetiology in complex diseases.

Even with these new findings, in a field with a modest quantity of replicated genetic associations, the proportion of heritability accounted for by copy number variants is likely to be low. Genetic association studies are currently under-powered to detect interaction effects and the effects of variants with a small effect size, although large consortiums of collaborating research centres may help with this problem(Cichon et al., 2009). Untangling the interacting effects of different genetic variants is likely to be particularly tricky, but could pave the way for a new aetiological model for susceptibility to psychiatric disorders.

Chapter 2. Rare Copy Number Variants



Parts of this analysis have been published in Rucker, J. J. H., Breen, G., Pinto, D., Pedroso, I., Lewis, C. M., Cohen-Woods, S., Uher, R., et al. (2011). Genome-wide association analysis of copy number variation in recurrent depressive disorder. Mol Psychiatry, –. Nature Publishing Group. doi:10.1038/mp.2011.144

2.1 Introduction

In this chapter I will describe the workflow for the analysis of rare copy number variants (CNVs) in this project, followed by the results of this analysis. As seen in the previous chapter, rare CNVs have been associated with a number of neuropsychiatric disorders, and it is therefore a novel research question to investigate this in clinically ascertained recurrent depressive disorder (RDD), which may represent a more heritable form of the disease than non-recurrent depression or depression ascertained via population surveys(McGuffin, Cohen, & Knight, 2007).

2.2 Hypotheses

We hypothesise that

A) We will find an increased frequency of samples with rare CNVs in our case sample.

B) We will find an increased frequency of CNVs in loci previously implicated in other psychiatric disorders

C) We will find increased numbers of singleton events in RDD cases.

Counterbalancing our hypotheses, we should note that there is conflicting evidence that rare CNVs are associated with the other major psychiatric affective disorder, bipolar disorder (see, for example(Grozeva et al., 2010; Malhotra et al., 2011; McQuillin et al., 2011; Priebe et al., 2011)).

2.3 Methods

2.3.1 Samples

SUMMARY: This larger study draws its cases and controls from three existing cohorts of patients with recurrent unipolar depression, a further screened control cohort from a study of bipolar affective disorder and a further cohort of population controls from phase two of the Wellcome Trust Case Control Consortium. The mean age of sampling for our cases was 45.2 years (SD 12.2 years) with a mean age of onset of 26.1 years (SD 12.0 years). 1,692 cases (54.5%) were of non-UK European heritage. Screened control samples had a mean age of 41.4 years (SD 13.2 years) and were exclusively of UK origin.

2.3.1.1 The Depression Case Control Study

The Depression Case Control (DeCC) sample consists of 1,536 cases (69% women) of recurrent ICD10(World Health Organisation, 1993)/DSMIV(American Psychiatric Association, 1994) criteria diagnosed and SCAN (Schedules for Clinical Assessment in Neuropsychiatry(Wing et al., 1990)) assessed recurrent unipolar depressive disorder of at least moderate severity. Subjects were from three clinical UK sites in London, Cardiff and Birmingham, drawn from psychiatric clinics, hospitals and general medical practices and from volunteers responding to media advertisements. Subjects were excluded if they, or a first-degree relative, had a history of mania, hypomania, or schizophrenia. Subjects were also excluded if they experienced mood incongruent psychotic symptoms or had histories of intravenous drug use with a lifetime diagnosis of dependency or depression occurring solely consequent to substance misuse or

medical problems. Mean age at recruitment was 47.5 years, SD 11.3 years, range 19-85 years.

1,500 controls (60% women) were drawn from the MRC general practice research framework and healthy volunteers drawn from staff at King's College London. Controls were screened with the composite index of depressive and anxiety symptoms and the Past History Schedule (Gottesman, 1991; McGuffin, Katz, & Aldrich, 1986). Mean age at recruitment was 47.9 years, SD 9.1 years, range 22-68 years. Only subjects with exclusively white European parentage were included, identified by self-report. Venous blood samples were collected from cases, whilst cheek swab samples were collected from controls. Further details on this cohort can be found in Gaysina et al.(Gaysina et al., 2008).

2.3.1.2 The Depression Network Study

1,200 probands (76% women) from an affected sib-pair linkage study called the Depression Network Study (DeNt) were included as part of this study. Individuals were recruited from psychiatric clinics, hospitals, general medical practices and from volunteers responding to media advertisements in 8 clinical sites around Europe. The probands selected were aged between 18-65 years old and consisted of recurrent ICD10/DSMIV criteria diagnosed and SCAN assessed recurrent cases of unipolar depressive disorder of at least moderate severity with identical phenotypic assessment to the DeCC sample. Exclusion criteria were identical to DeCC. Mean age at recruitment was 45.1 years, SD=11.8 years, range 18-78 years. All probands were of exclusively white European parentage, identified by self-report. Venous blood samples were collected from all

participants. Further details on this cohort can be found in Farmer et al.(Farmer et al., 2004).

2.3.1.3 The Genome Based Therapeutic Drugs for Depression (GENDEP) Study

The Genome Based Therapeutic Drugs for Depression (GENDEP) study, was a pharmacogenetic investigation into antidepressant response in individuals with depressive disorders. GENDEP included 868 treatment-seeking adults (63% women) diagnosed with ICD-10/DSM-IV unipolar major depression of at least moderate severity established via interview with the SCAN. Mean age at recruitment was 42.3 years, SD 11.5 years, range 19-72 years. Eligible participants were treated either with the predominantly noradrenergic drug Nortriptyline or the highly selective serotonergic drug Escitalopram over a 26-week period. Participants were drawn from referrals from general practice and psychiatric services and by advertisement in nine European centre. Only those with white European ethnicity were included (assessed by self-report). Potential participants were excluded if they had a family history of bipolar affective disorder or schizophrenia in a first-degree relative, a personal history of hypomanic or manic episodes, schizophrenia, mood incongruent psychotic symptoms, primary substance misuse, primary organic disease and pregnancy.

The only phenotypic difference from DeCC and DeNt is that GENDEP patients need not have suffered from recurrent disorder (although, as it turned out, the majority, 61%, had recurrent disorder)(Uher et al., 2009). All samples were included in this analysis. Venous blood samples were collected from all participants.

2.3.1.4 The Bipolar Case Control Study

Further screened control samples are derived from the Bipolar Case Control study (BACCs), comprising 459 controls (61% women). Potential subjects for the screened control group were collected from students and staff at Kings College London by internal email advertisement and by local media advertisement. Subjects were interviewed with a modified version of the Past History Schedule (McGuffin et al., 1986) and were included only if they demonstrated no evidence of past or present psychiatric disorder. Subjects were also interviewed with the Beck Depression Inventory (BDI) and excluded if they scored greater than 10. Mean age at recruitment was 32.5 years, SD=12.5 years, range 18-89 years. All participants were of white European heritage (assessed by self-report). Venous blood samples were collected from all participants. More detail on this cohort can be found in Gaysina et al.(Gaysina et al., 2009).

2.3.1.5 The Wellcome Trust Case Control Consortium (Phase 2)

As an additional control set we also used 5,619 control samples (46% women) from phase 2 of the Wellcome Trust Case Control Consortium (WTCCC2), which is composed of the 1958 British birth cohort and the national blood service cohort. Both cohorts represent population control samples and therefore members are not excluded based on a diagnosis with a particular disease. Further details on these cohorts can also be found in Craddock et al.(Craddock et al., 2010).

2.3.1.5.1 The 1958 British Birth Cohort

The 1958 British birth cohort (also known as the National Child Development Study) is taken from participants in a longitudinal study derived from live sequential births in the UK (England, Wales and Scotland) during one week of 1958. The cohort was followed up in 2002-4, making the participants 44-46 years of age("Genetic SNP information from the 1958 British Birth Cohort," n.d.). DNA samples from the 1958 British birth cohort are derived from cell lines immortalised via transformation with Epstein Barr virus.

2.3.1.5.2 The National Blood Service Cohort

The national blood service cohort is derived from volunteers donating blood to the UK blood collection service, with an age range of 18-69 years. Individuals from the national blood service cohort have been screened by standard processes used to exclude blood donors from groups unsuitable to donate blood. The cohort represents individuals selected to mimic the geographical distribution of the 1958 birth cohort. DNA samples derived from the national blood service cohort are derived from venous blood.

2.3.2 Genotyping

SUMMARY: Our samples were genotyped on Illumina 610 Quad and modified 1M Infinium beadarrays. Our cases and screened controls were genotyped contemporaneously at the same laboratory however the WTCCC2 controls were genotyped elsewhere. We used standard Illumina methodology to cluster markers and relative fluorescence intensity values and relative allelic intensity values. We used PennCNV to make copy number calls using a consensus marker set between the two Illumina arrays.

2.3.2.1 Introduction

Genotyping was performed on Illumina beadarrays(Gunderson et al., 2004) with infinium genotyping assays(Steemers et al., 2006). The cases and screened controls were genotyped on the Illumina 610 Quad beadchip whilst the population control sample from the WTCCC2 was genotyped on a modified version of the Illumina 1M beadchip. Both arrays use an identical infinium assay and identical Illumina beadarray technology. Briefly, beadarray technology is based on barcoded 3-micron silica beads that self-assemble in microwells on silica slides. To each bead is attached hundreds of thousands of copies of a specific oligonucleotide sequence unique to the reference genome and representing either a single nucleotide polymorphism (SNP), which by definition is biallelic, or a monoallelic probe designed purely to measure relative copy number. Infinium assays are based on whole-array hybridization of whole genome amplified DNA samples to unique probes attached to the silica beads described above. A single base extension and amplification step of two

fluorescently labelled markers representing the two alleles of each SNP then allows each barcoded bead to be read by a laser. For further details, two original papers describe beadarray technology(Gunderson, Steemers, Lee, Mendoza, & Chee, 2005) and the infinium assay(Steemers et al., 2006) comprehensively.

2.3.2.2 Case and Screened Control Samples

Venous blood was collected from participants at the time of interview in 7.5ml EDTA monovette tubes, shaken gently and stored at -20C prior to DNA extraction. DNA from the DeNT screened control population was derived from cheek swabs collected either at the time of interview, or via kits delivered by mail. DNA from the DeNT sample was extracted using columns (DNeasy Blood & Tissue Kit, Qiagen, UK), whilst DNA from DeCC, GENDEP and BaCC samples were extracted using the phenochloroform method.

Concentration, fragmentation and response to PCR were determined in all samples. Samples with poor DNA quality were either re-extracted or excluded. Aliquots of 15µl of DNA at 50ng/µl were prepared and robotically dispensed into barcoded 96 well plates. Samples were randomly distributed amongst the plates. Genotyping was performed at the Centre Nationale De Genotypage (CNG) in Evry, near Paris, France. The CNG is a fully automated Illumina BeadLab equipped with liquid handling robots (Tecan Ltd, Dorset, UK), Illumina BeadArray readers and Illumina iScans (Illumina Ltd, CA, USA).

All probe intensity data was normalised and processed by the CNG using standard Illumina protocols with Illumina's GenomeStudio platform to obtain the log R ratio (LRR) and B allele frequency (BAF) at each marker. The LRR and BAF

represent, for each marker in each sample, a summed probe intensity ratio derived from comparison to a canonical value calculated from all samples, and, in the case of biallelic probes, an allelic intensity ratio, respectively. The CNG sent us Illumina 'Final Report' files consisting of SNP allele calls, LRR and BAF values and raw and normalised X and Y values, representing the intensity of the fluorescent dyes CY3 and CY5 used for each marker.

2.3.2.3 Population Control Samples (WTCCC2)

2.3.2.4 Derivation of Normalised Probe Intensity Data from GenomeStudio

We applied for, and were granted access to, the Illumina 1M 'idat' files for the control samples of phase 2 of the Wellcome Trust Case Control Consortium, including the 1958 birth cohort and the national blood service cohort. Idat files are generated by the Illumina iScan device when microarray chips are scanned after sample hybridisation and represent the raw intensity data derived from each probe on the array. Such files may be imported in Illumina's proprietary 'GenomeStudio' platform, to form a GenomeStudio 'project'. An initial experiment-wide normalization process which takes account of DNA concentration differences and plate effects is automatically applied to data from idat files as they are imported into a new project (Peiffer et al., 2006). Marker by marker normalisation and transformation is then performed in a two-step process illustrated for one SNP marker (rs12414155) in Fig 2.1.

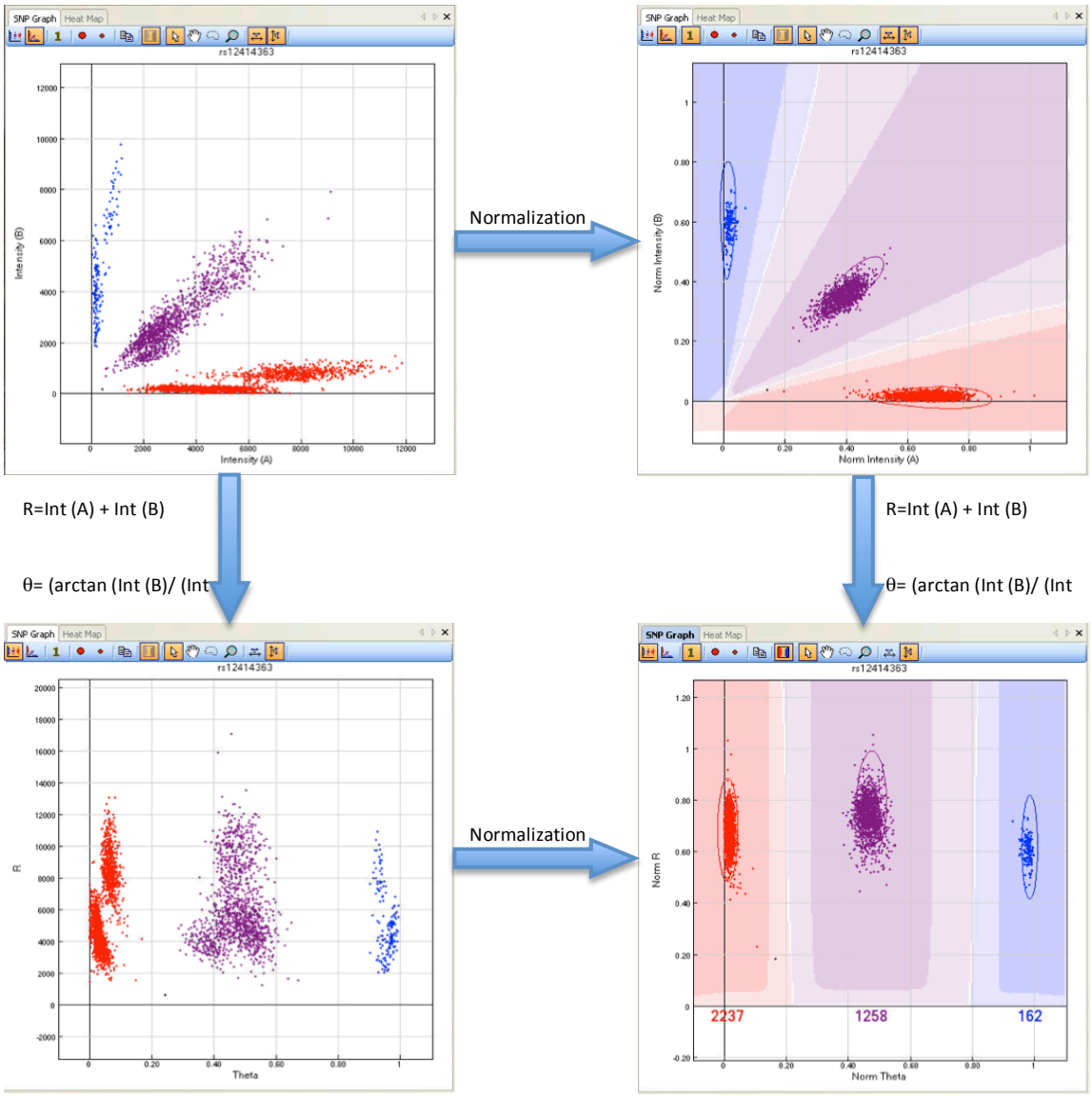


Fig 2.1. Normalization (left to right) and transformation (top to bottom) of raw intensity values for one SNP (rs12414155).

The initial data plot for one marker is seen in the upper left corner of Fig. 2.1, and represents the raw intensity values for each fluorescent dye representing each allele of the marker (which Illumina arbitrarily denote as allele 'A' and allele 'B'). The process of inter-sample normalization (left to right images) and data transformation (top to bottom images) within GenomeStudio is used to derive two data points. 'R' reflects the sum of the probe intensities and 'θ' reflects the arctangent of the allelic intensity ratio B/A, divided by $\pi/2$ (bottom right image).

This maximises the separation of the three clusters of a standard bi-allelic marker, which eases the subsequent process of calling genotypes. Normalisation and transformation of data is not dependent on the sequence in which it occurs.

Without a reference value for each allele at each marker one cannot make any analysis of the relative copy number state. Thus the next step in processing intensity data is to derive a reference value for each marker. Illumina provides two methods to derive a reference value. The first, least preferred method, is to use a reference file supplied by Illumina for the chip type being used, where the reference values are derived from HapMap samples run on the same chip. This is the less preferred method because differing conditions between laboratories are likely to mean that systematic differences exist from the Illumina supplied reference file and the user's sample data. The second, preferred method, is to derive a reference value from the samples within one's own project, provided there are a sufficient number of samples. Within our analyses we chose the latter, as there are more than sufficient samples to derive a robust reference value for each marker.

The GenTrain algorithm within Illumina's GenomeStudio application is then called to automatically cluster, call genotypes and assign confidence scores to individual markers. We filtered our samples and clustered our data according to Illumina's recommended protocol. Briefly, the GenTrain algorithm will attempt to cluster and assign a score to each SNP marker. Some SNPs will behave poorly or fail clustering, and these can be identified by various metrics calculated by GenomeStudio and then need to be manually reclustered or zeroed (excluded).

SNPs in regions of increased copy number often have indistinct clusters representing the increased number of possible genotypes. Allelic clustering can fail in these SNPs, however a relative intensity ratio can still be calculated from the sum of the probe intensities (R) compared to the reference derived from all samples. We evaluated by eye all markers with a GenTrain score of less than 0.6 and manually re-clustered markers as per Illumina's protocol. Markers that could not be reclustered were zeroed.

After a reference value for each marker has been calculated, and genotypes assigned, GenomeStudio calculates two measures which are used for the onward derivation of copy number calls. The Log R Ratio (LRR) is calculated as a ratio measure of overall normalized signal intensity for each marker where $LRR=\log_2(R_{observed}-R_{expected})$. The B allele frequency (BAF) is a measure of the relative signal intensity ratio of each allele, A and B:

$$BAF=\begin{cases} 0, if & \theta < \theta_{AA} \\ 0.5\dfrac{(\theta-\theta_{AA})}{(\theta_{AB}-\theta_{AA})}, if & \theta_{AA} \le \theta < \theta_{AB} \\ 0.5+0.5\dfrac{(\theta-\theta_{AB})}{(\theta_{BB}-\theta_{AB})}, if & \theta_{AB} \le \theta < \theta_{BB} \\ 1, if & \theta \ge \theta_{BB} \end{cases}$$

where θ_{AA}, θ_{AB} and θ_{BB} are the θ values generated from many canonical samples based on an assumption of a biallelic genotype. The LRR will generally vary from a lower value of about -6, representing a full, or homozygous, deletion, to a value of about 1, representing a duplication of ≥4 copies. The B allele frequency varies between 0 and 1 with clusters around 0, 0.5 and 1 for biallelic genotypes.

After automatic and manual clustering, we then exported Final Report files from GenomeStudio, which are text files containing, amongst other data, the LRR and BAF for each marker in each sample. We also exported genotype call rates as a quality control metric for our samples.

2.3.2.5 Copy Number Calling

We used PennCNV to make copy number calls. PennCNV is a popular, open-source package designed for Illumina array data that implements a hidden Markov model, a Viterbi algorithm, expectation maximisation and takes account of the distance between consecutive markers to make copy number calls (K. Wang et al., 2007). All three of these techniques derive from probability theory.

2.3.2.5.1 Hidden Markov Models

A hidden Markov model (HMM) attempts to represent a non-deterministic (adaptive) model of behaviour composed of a finite number of 'states', transitions between those states, and actions. For the purposes of CNV detection the states represent copy number states. Non-deterministic HMMs produce probabilities of emission states and transition states, rather than the states themselves. The emission state represents the probability of the marker being in a specific copy number state, whilst the transition probability describes the probability of having a change of state between one marker and the next. This is reliant on the genomic distance between markers, which PennCNV takes into account. This is modeled in PennCNV by a state transition matrix that was previously described by Marioni (Marioni, Thorne, & Tavare, 2006).

2.3.2.5.2 Expectation Maximisation

Expectation maximisation is a natural generalisation of maximum likelihood estimation, which is a method of estimating the parameters of any given statistical model, extended to the estimation of unknown or incomplete data(Dempster, Laird, & Rubin, 1977). In the context of CNV calling, expectation maximisation is used to update the prior parameters of the HMM upon which a CNV call may be modelled, those prior parameters themselves being estimated from observed data (in PennCNV, from plots of large CNV regions observed using GenomeStudio with HumanHap 550k arrays).

2.3.2.5.3 Viterbi Algorithms

The Viterbi algorithm is a method for calculating the most likely path (known as the Viterbi path) from a series of hidden event states(Viterbi, 1967). The algorithm makes three assumptions, which happen to fit microarray data well. Firstly the events under observation must occur in sequence, secondly they must be aligned and thirdly the most likely path through the events is a reflection of the observed end point and all the event states leading up to it. In this sense, the Viterbi algorithm takes the emission probabilities for each sequential marker on the array from the hidden Markov model, and calculates the most likely path (or sequence of copy number states) given those probabilities.

Finally a CNV is called by PennCNV when a stretch of markers (>2 by default, but this can be set by the user) are given states that deviate from the normal state (copy number of 2).

2.3.2.5.4 PennCNV Calling

Having obtained final report files from our cases and screened control cohorts and the WTCCC2 control cohorts, we split our Illumina final report files into files representing the LRR and BAF calls for each sample individually. To do this we modified a Perl script included with the PennCNV package to produce files with chromosome, position, marker name, LRR and BAF values. This was done in order that these files could be used for functions other than calling CNVs with PennCNV.

Because the WTCCC2 cohort was genotyped with the Illumina 1M chip, and our cases and screened control samples were run with the Illumina 610 Quad chip, we restricted our data to a common set of markers between the two arrays. This consisted of 562,680 markers.

We called our CNVs using PennCNV after liaison with the author, Kai Wang at Penn State University, to determine which hidden markov model parameters and pfb (population frequency of the B allele) file to use with our data. We used the prior parameters file for the hidden Markov model which allows non-SNP markers to be processed (hhall.hmm) and the hhall.pfb file, the values for which are "compiled from a large set of individuals with mixed ethnic backgrounds and of normal phenotype"(K. Wang et al., 2007). The command settings for PennCNV can be seen in the appendix. We processed a total of 10,457 samples, summarized in table 2.1, to produce copy number calls and sample QC statistics for onward processing and analysis.

Cohort	Sub Cohort	Description of Samples
Depression GWAS	Depression Network study	1,092 cases
	Depression Case Control study	1,268 cases and 1,272 screened controls
	Genome Based Therapeutic Drugs for Depression	746 cases
	Bipolar Case Control sample	459 screened controls
Wellcome Trust Case Control Consortium, phase 2	1958 British birth cohort	2,920 controls
	National blood service cohort	2,699 controls

Table 2.1. Description of Samples used in CNV analysis.

2.3.3 Visualisation of Data

SUMMARY: We visualised our samples and calls using scripts developed within the R programming environment and the functionality provided by the PennCNV program.

A pressing problem within our analysis was the visualisation of LRR/BAF values by sample, and by call. Whilst we could use the functionality within GenomeStudio to visualise samples from the WTCCC2, we could not use a similar method to visualise samples from our cases and screened controls because we had not been provided with the idat files. Thus we developed a script written in the programming language R(R Development Core Team, 2005), which can be seen in the appendix, to plot the LRR and BAF values by chromosome and by sample. We also used a package provided within the PennCNV program to plot calls made by the program. We also wrote our own script to plot normalised R and theta values by marker, which can also be viewed in the appendix.

2.3.4 Sample Quality Control

SUMMARY: We attempted to strike a balance between type 1 and type 2 errors in our quality control of samples and calls, which is a pressing and tricky problem in CNV analysis. We used the GCR, BAFSD, LRRSD and total number of calls to set exclusion thresholds for samples, and a variety of criteria to exclude calls that were likely to be artefactual or irrelevant to our analysis. We found that samples derived from cheek swab DNA gave systematically different results to samples derived from blood cells and excluded them from our analysis.

Quality control (QC) of samples and calls in CNV analysis can be summarised as an attempt to strike a balance between making type 1 errors (retaining false positive calls) and type 2 errors (excluding true calls). This larger aim may be potentially achieved by three methods

1. Removing samples with QC metrics that suggest a higher probability of making false positive calls.
2. Removing individual calls that are more likely to be false positive.
3. Visually QCing samples and calls to remove those likely to be false positive.

The volume of data in this analysis precludes any sort of systematic visual QC of samples or calls, which in any event may be subjectively biased. Similarly it is almost impossible to prevent type 2 errors if the calling algorithms do not make the calls in the first place. Thus we relied on strategies 1 and 2 to exclude false positive calls.

We initially analysed the quality of our samples based on three metrics, the genotype call rate (GCR) from GenomeStudio, the standard deviation of the Log R Ratio (LRRSD) and the standard deviation of the B Allele Frequency (BAFSD). These, broadly speaking, are measures both of initial DNA quality at the time of running samples on the array, and of the array experiment itself. Higher LRRSD and BAFSD values and low GCR values generally indicate samples of poorer quality which tend to produce a higher proportion of false positive calls.

2.3.4.1 Cheek Swab DNA

We were mindful of the fact that the DNA samples from screened controls collected as part of the DeCC study were derived from cheek swabs, rather than venous blood. Indeed, many of these samples were collected by mailing participants DNA collection kits in the post, meaning supervision of sample collection could not be carried out, and appropriate technique therefore not guaranteed. We decided to perform an investigation of our quality control metrics stratified by DNA source, to check for systematic differences between samples derived from venous blood and samples derived from cheek swabs.

Table 2.2 shows the mean and standard deviation for the LRRSD, BAFSD and GCR stratified into samples derived from venous blood and cheek swabs. For all three metrics, and especially LRRSD and BAFSD, cheek swab DNA samples show a less favourable mean and standard deviation, suggesting overall that cheek swab samples are generally of lower quality than venous blood samples. Figs 2.2, 2.3 and 2.4 illustrate boxplots for LRRSD, BAFSD and GCR respectively, stratified by

DNA source. These plots illustrate the systematic differences in quality between the two DNA sources.

DNA Source	Variable	Mean	SD	Min	Max
Blood	LRRSD	0.187	0.069	0.090	1.039
Chk Swab		0.301	0.120	0.117	0.700
Blood	BAFSD	0.035	0.009	0.021	0.146
Chk Swab		0.045	0.011	0.027	0.091
Blood	GCR	0.997	0.008	0.825	1.000
Chk Swab		0.992	0.014	0.905	0.999

Table 2.2. Summary statistics for LRRSD, BAFSD and GCR, stratified by DNA source across cases and screened controls.

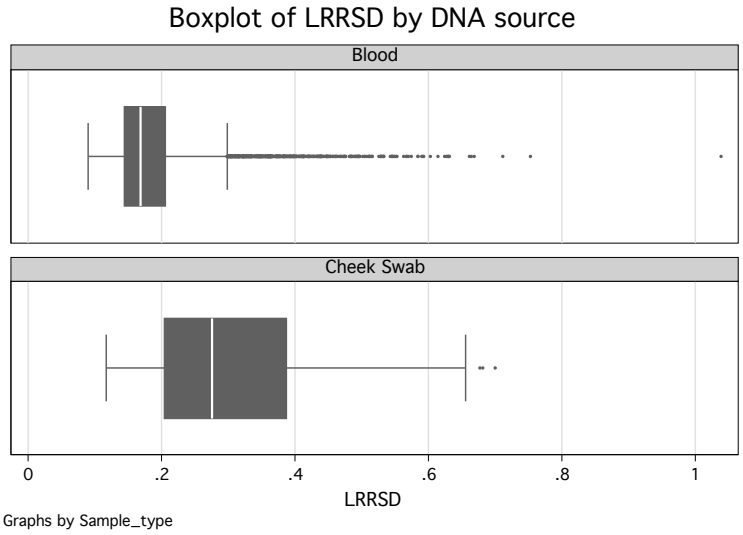


Fig 2.2. Boxplot of LRRSD, stratified by DNA source.

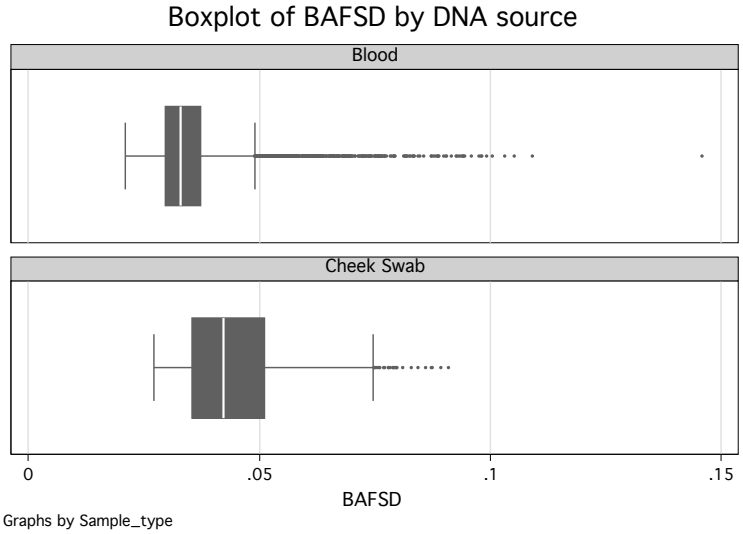


Fig 2.3. Boxplot of BAFSD, stratified by DNA source.

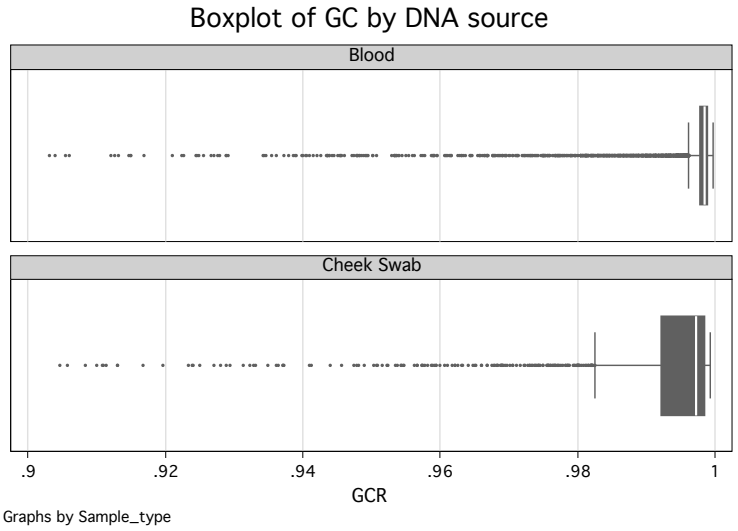


Fig 2.4. Boxplot of genotype call rate (GC), stratified by DNA source.

Given the differences in sample QC metrics indicated above, we next compared the number of CNV calls made in autosomes by PennCNV in the two DNA source groups.

Table 2.3 illustrates summary statistics for the number of autosomal CNVs called by PennCNV stratified by DNA source, whilst Fig 2.5 illustrates this data in a boxplot.

DNA Source	Variable	Mean	SD	Min	Max
Blood	No. Autosomal CNVs	20.97	27.85	1	248
Chk Swab		29.11	24.59	4	237

Table 2.3. Number of autosomal CNVs called by PennCNV in samples derived from venous blood and cheek swab DNA.

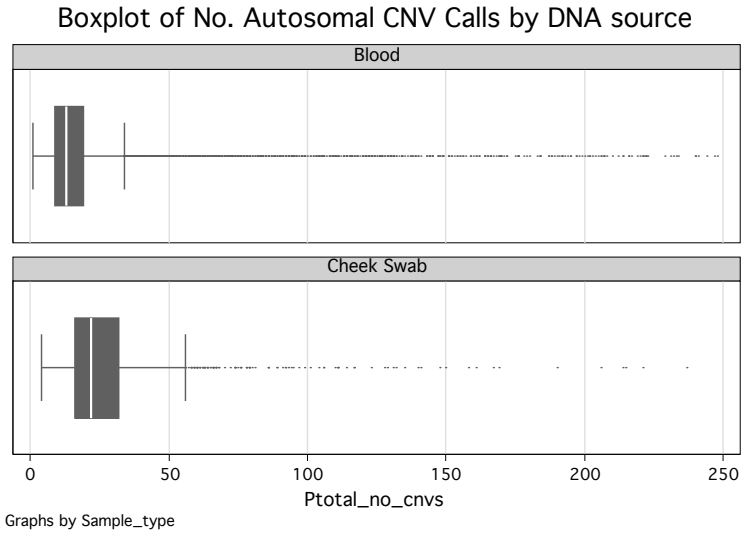


Fig 2.5. Number of autosomal CNVs called by PennCNV, stratified by DNA source.

To formally test whether, in terms of the number of CNVs called, cheek swab samples were systematically different to blood samples we performed a two-sample Wilcoxon rank-sum (Mann-Whitney) test. We found that cheek swab samples were indeed significantly different ($z=-28.9$, $p>|z| <0.00001$). We visually examined the types of CNV found in our cheek swab samples and found that the increased number of CNVs was driven by a variety of apparently spurious calls, including an excess of large (>100kb) duplication events. These

occurred in a non-predictable manner, and were not confined to samples with poor QC metrics. They may indicate contamination with other sources of DNA at the time of sampling, since many of the cheek swab samples were done by participants in their own home, and not supervised by a researcher. Given this observation we decided to exclude our cheek swab samples from our analysis of rare CNVs. In practice this means excluding the screened control samples taken as part of the DeCC study, but retaining those taken as part of the BaCCs study (which were all derived from venous blood samples).

2.3.4.2 Analysis of QC Metrics by Sub-Cohort

After excluding our cheek swab samples, we plotted the LRRSD, BAFSD and GCR of our remaining samples, stratified by sub-cohort, to gauge the distribution of QC metrics within the sub-cohort of our sample groups and to provide a basis for selecting QC metric cut offs that both maximised sample numbers whilst removing samples of poor quality. Figs 2.6, 2.7 & 2.8 illustrate histograms, stratified by sub-cohort, of LRRSD, BAFSD and GCR.

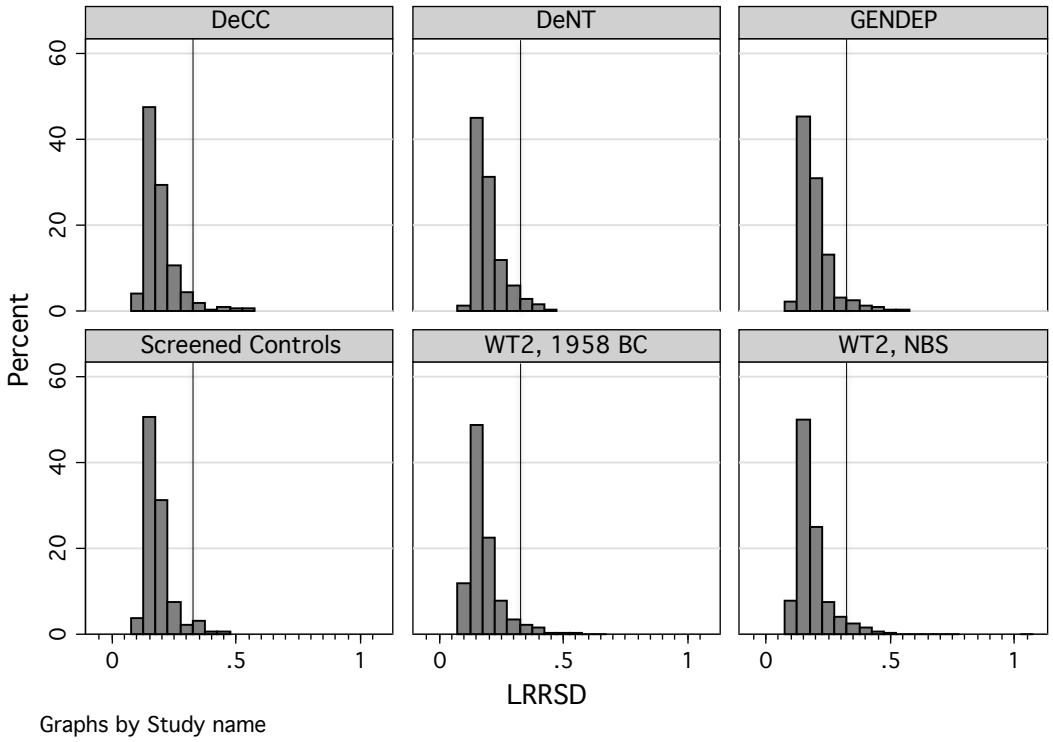


Fig. 2.6. Histogram of log R ratio standard deviation (LRRSD) by sample sub cohort. Vertical line indicates QC cut off.

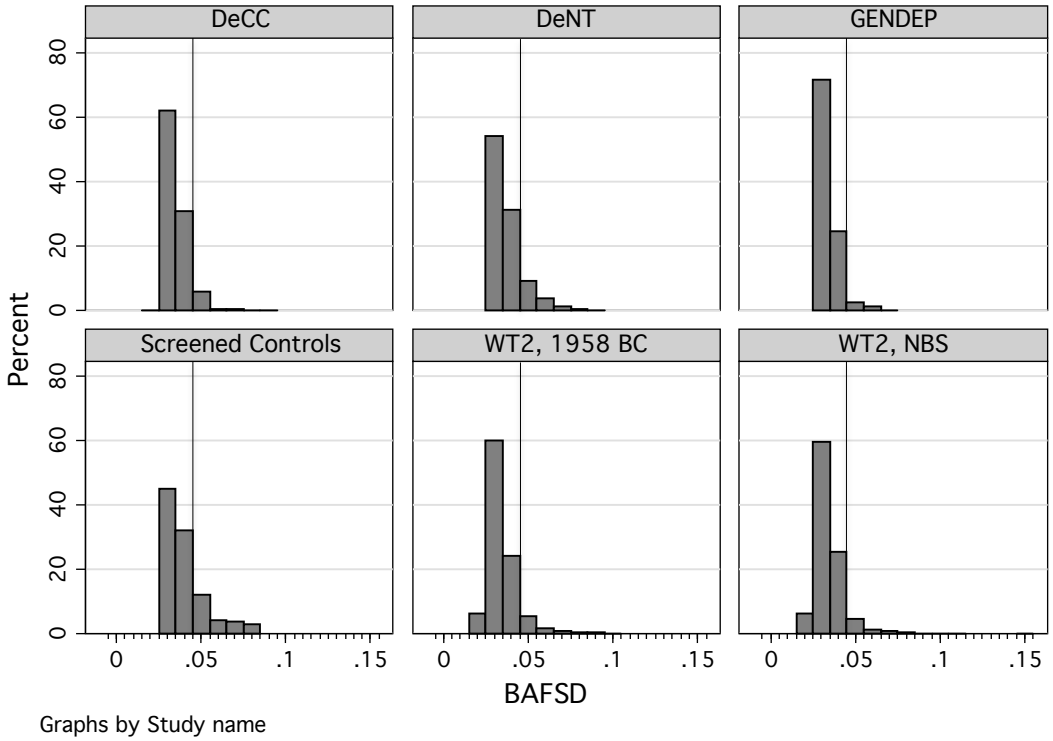


Fig. 2.7. Histogram of B allele frequency standard deviation (BAFSD) by sample sub cohort. Vertical line indicates QC cut off.

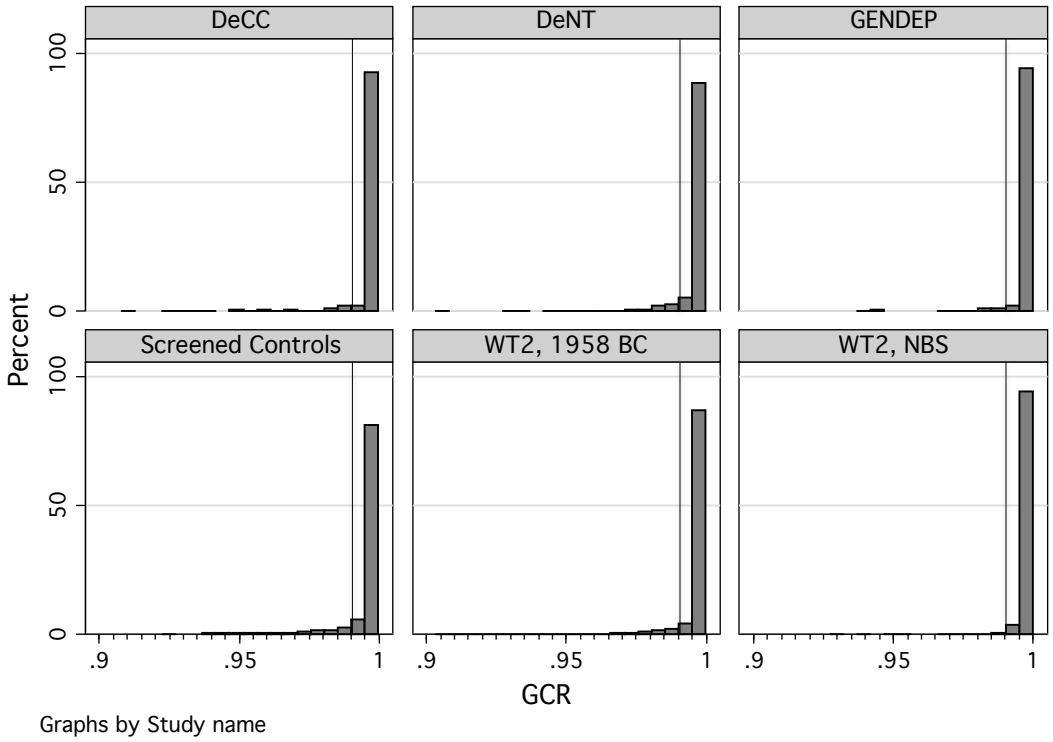


Fig. 2.8. Histogram of genotype call rate (GCR) by sample sub cohort. Vertical line indicates QC cut off.

We set our sample quality control metrics, excluding initially any sample that failed any one of the following three limits

1. GCR < 98%
2. LRRSD > 0.3
3. BAFSD > 0.045

A plot of LRR and BAF for markers in chromosome 1 in samples at the lower end of our QC thresholds by LRRSD, BAFSD and GCR is presented in Figs. 2.9, 2.10 and 2.11. In particular these plots demonstrate waviness (fluctuating deviation from the median) for LRR values (red plots) and noisy (non-fluctuating deviation from the median) for LRR and BAF values (blue plots).

Whilst the metrics described so far do exclude a majority of samples liable to generate false positive CNV calls (type 1 errors), we also describe steps in the next section to further refine the sample set for this analysis.

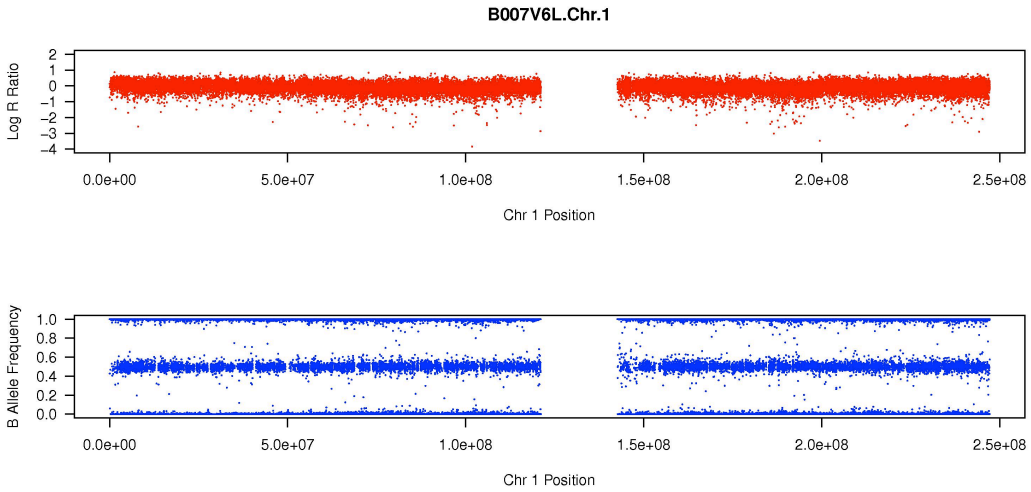


Fig 2.9. LRR/BAF values for chromosome 1 in sample B007V6L, with an LRRSD of 0.3. Some waviness and extreme values can be seen within the plot of LRR values (upper plot)

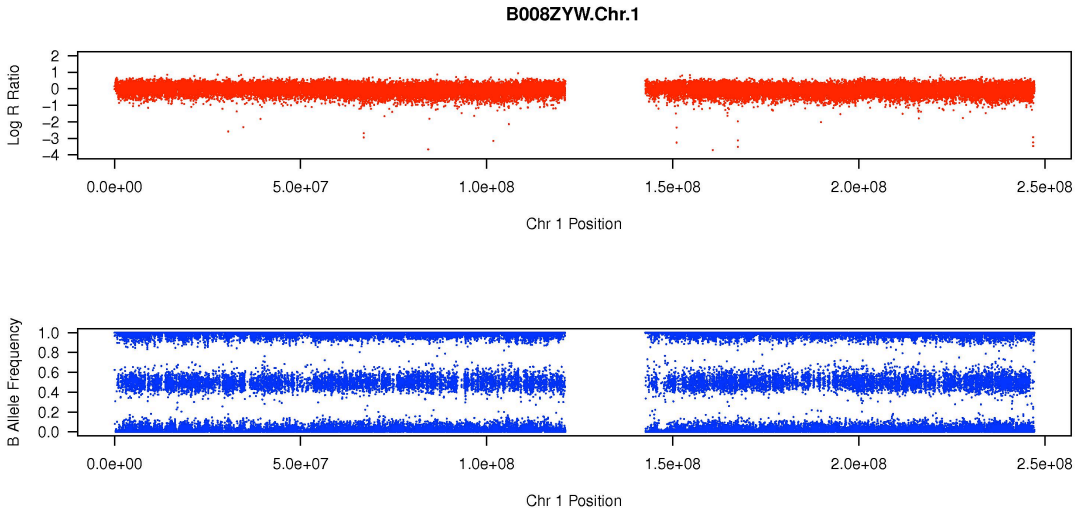


Fig 2.10. LRR/BAF values for chromosome 1 in sample B008ZYW, with a BAFSD of 0.045. BAF values have a tendency to deviate from the median value of 0.5 (lower plot)

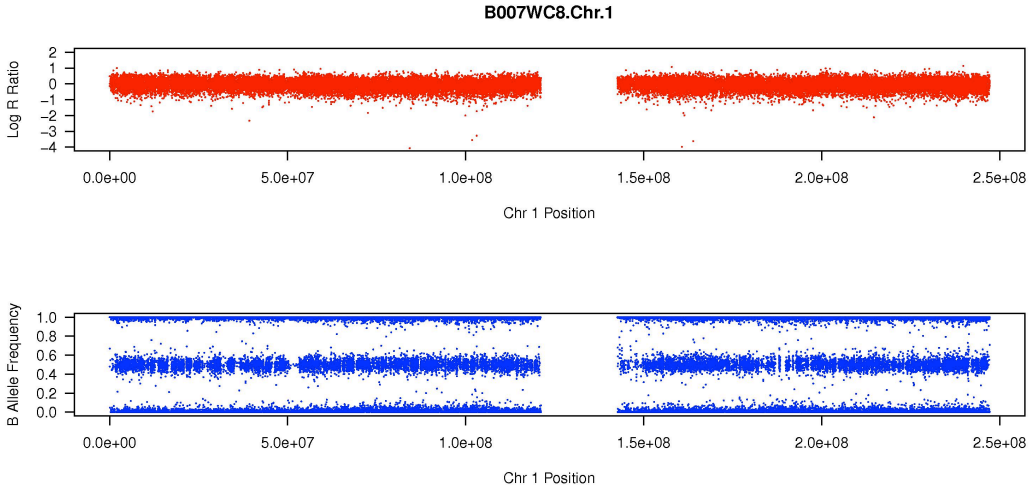


Fig 2.11. LRR/BAF values for chromosome 1 in sample B007WC8, with a GCR of 0.98. This sample shows minor waviness in the LRR plot (upper plot), and BAF values deviating from the median (lower plot).

2.3.5 CNV Quality Control

We based our quality control of individual CNV calls on three metrics.

1. The size of the call
2. The number of sequential markers used to make the call
3. The genomic location of the call

Small CNVs and CNVs called with few markers are more likely to include false positive calls secondary to random noise and genomic artefact. Some regions of the genome are particularly prone to degradation, such as the telomeres and centromeres. The immunoglobulin regions are particularly prone to rearrangement in lymphoblastoid cells(Lieber, Yu, & Raghavan, 2006).

For our analysis of rare CNVs we excluded calls that:

1. Were <100kb in length or were made with <10 consecutive markers

2. Occurred in regions within 500kb of the centromere or telomere or in immunoglobulin regions (see appendix for precise coordinates)
3. Occurred in regions where the total number of CNVs in the sample was greater than 1% of the total sample number (defined as a common CNV, or CNP)
4. Occurred in regions where the marker density was less than 1 marker per 200,000bp.

Scripts for steps 1-4 are contained in the appendix.

2.3.5.1 Samples Passing QC with High Numbers of CNV Calls

We observed that exclusion criteria based on the LRRSD, BAFSD and GCR do not exclude a subset of samples with large numbers of CNV calls due to genome-wide and chromosome specific artefacts as well as rarer aneuploidy and mosaicism. As other studies have done (see, for example(McQuillin et al., 2011; N. M. Williams et al., 2010; D. Zhang et al., 2009a)) we excluded samples based on total number of CNVs. The selection of a limit is dependent on the type of microarray used and analysis platform, although for our analysis, because we used a consensus marker set and the two chip types (Illumina 610 and 1M) used a similar assay and underlying technology, we were able to set a cross-cohort limit. We plotted box plots (Fig. 2.12) to determine the appropriate level of cut off in our samples.

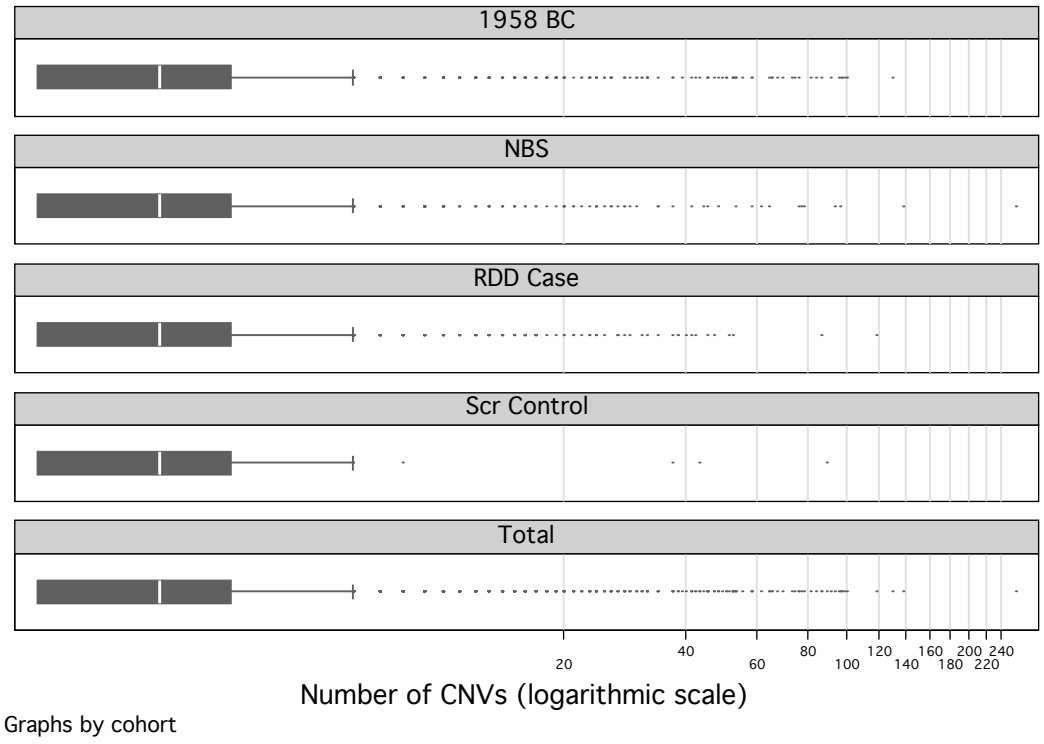


Fig. 2.12. Boxplot of number of CNVs per sample, stratified by cohort, after restriction by size and genomic location.

The plots indicate a similar distribution of samples (note that the screened control group is 6-fold smaller than the case and population control cohorts), with the vast majority of samples having between 0 and 10 calls, with a minority (108, 1.7%) exceeding 20. On visual examination these samples had a higher frequency of abnormalities such as waviness, and chromosomal aneuploidies attributable to cell-line transformation. Whilst these observations are in themselves interesting, they are not pertinent to our current analysis, and we excluded any sample with a total CNV count of over 20 and replotted histograms of number of CNVs per sample by sub-cohort (Fig. 2.13).

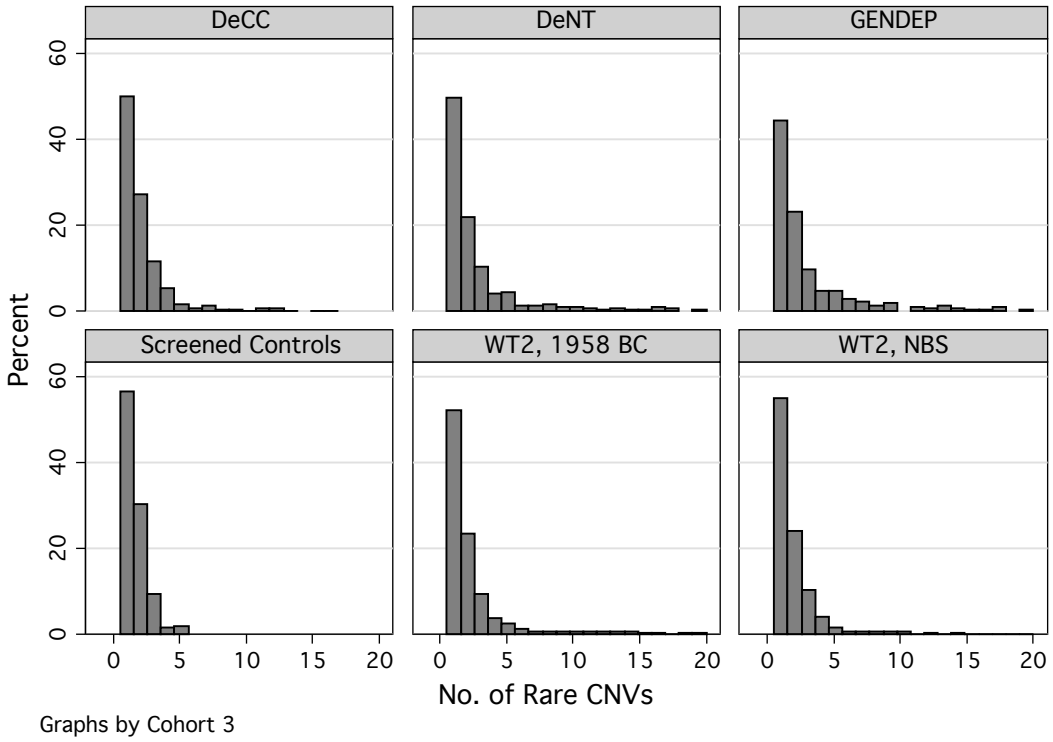


Fig. 2.13 Number of rare CNVs per sample, presented by sub-cohort.

2.3.6 Details of Samples and Calls Included

2.3.6.1 Samples

Our QC strategy can be divided into two stages, firstly the exclusion of samples via genotype call rate (GCR) calculated from GenomeStudio, and secondly the exclusion of samples based on LRRSD, BAFSD and number of CNV calls calculated by PennCNV. We present a table of the proportion of samples excluded at each stage in table 2.4. We noted that in our screened control sample a significant number of samples were excluded at stage 1. This was traced back to poor sampling technique by one research assistant at the time that blood samples were collected resulting in a batch of samples with poor DNA quality.

QC stage	Gender	Cases	Screened controls	WTCCC2 controls		
				NBS	58C	Gender total
All samples	Male	100%	100%	100%	100%	100%
	Female	100%	100%	100%	100%	100%
	Total	100%	100%	100%	100%	100%
				100%		
Post QC stage 1	Male	98.15%	87.08%	95.46%	94.93%	95.18%
	Female	98.13%	95.73%	96.82%	95.85%	96.33%
	Total	98.13%	92.37%	96.15%	95.38%	95.75%
				95.75%		
Post QC stage 2	Male	90.68%	74.19%	90.50%	87.90%	89.14%
	Female	88.78%	86.62%	90.16%	90.54%	90.35%
	Total	89.34%	82.08%	90.33%	89.19%	89.74%
				89.74%		

Table 2.4. Sample quality control. Stage 1- exclusion by genotype call rate (<98%). Stage 2- exclusion by LRRSD (≤0.3) BAFSD (≤0.045) across autosomal chromosomes and total number of CNVs (≤20) across autosomal chromosomes. NB A larger proportion of screened control samples were excluded at QC stage 1 due to a batch of samples with poor DNA quality.

2.3.6.2 CNV Calls

Our CNV quality control can be subdivided into five stages. Numbers of variants and percentages of calls remaining as a proportion of the initial starting value are detailed in table 2.5.

Cohort/QC Stage	Cases (2,723)	Screened Controls (348)	WTCCC2 Controls (4,828)
All CNVs (post sample QC)	56,773	5,710	120,042
	100.00%	100.00%	100.00%
QC stage 1	7,492	736	34,408
	13.20%	12.89%	28.66%
QC stage 2	6,826	676	33,133
	12.02%	11.84%	27.60%
QC stage 3	5,953	507	9,288
	10.49%	8.88%	7.74%
QC stage 4	5,687	503	9,226
	10.02%	8.81%	7.69%
QC stage 5	4,285	321	6,790
	7.55%	5.62%	5.66%

Table 2.5. Numbers of CNVs at different quality control stages.

QC stage 1 Any CNV with a constituent number of SNPs of less than 10 or a length less than 100kb was removed.

QC stage 2 All CNVs with more than 50% overlap with immunoglobulin regions or regions defined as within 500kb of the centromere or telomere (see appendix) were removed.

QC stage 3 All samples with more than a total of 20 CNVs were removed.

QC stage 4 All CNVs from genomic regions with less than 1 marker per 200kb in our marker set were removed.

QC stage 5 All CNVs occurring at >1% frequency in our population control sample were removed from all cohorts.

The relatively large proportion of calls remaining in the WTCCC2 sample at QC stages 1 and 2 are likely to be attributed to variants introduced during cell line creation in the 1958 birth cohort.

2.3.7 Validation of CNVs

As part of a further in-depth study of the 22q11.2 region and 5 denovo CNV regions published by Stefansson et al.(Stefansson et al., 2008), we designed a custom oligonucleotide comparative genomic hybridisation (CGH) array (4x180k, Agilent Technologies, CA, USA) in liaison with Oxford Genome Technologies (Oxford, UK). The regions we analysed in our follow up analysis can be viewed in Table 2.6. 291 cases and 52 screened control samples were selected for follow up. DNA samples had already been extracted, cleaned and QC'd as part of the genotyping procedure for the SNP GWAS(Lewis et al., 2010). Nonetheless we verified DNA concentrations and 260:280 ratios before preparing 40µl of DNA at 50ng/µl in 96 well plates for transport to OGT.

Additional sample QC and normalisation were performed at the OGT laboratories. 1µg of sample DNA was hybridised against 1µg of Human Genomic Male DNA (Promega, WI, USA). The samples were labelled following Agilent's aCGH protocol 6.2.1. The purified labelled samples were dried down to completeness and then hybridised onto 4x180K arrays following the manufacturer's guidelines. Arrays were washed and scanned and feature-extracted as per the manufacturer's specifications with Agilent Feature Extraction 10.7.3.1. Raw data in the form of Log_2 ratios were sent back for processing.

We used OGT's supplied Cytosure software, which uses normalization and circular binary segmentation algorithms to normalize sample data, call and visualize copy number calls according to OGT's supplied protocol. We then visually followed up each call using the software's plots of Log_2 ratios and compared them to the calls made by PennCNV. We did not follow up X chromosome calls.

chr3:5,997,636-8,638,508
chr3:53,389,717-55,517,902
chr6:160,833,017-164,666,324
chr10:59,301,152-60,900,609
chr10:67,334,811-69,148,660
chr12:1,870,350-2,688,889
chr15:20,195,144-20,872,219
chr20:13,859,307-16,039,340
chr22:16,300,001-24,300,000
chrX: 5,390,291-8,806,740
Total coverage: 30,524,214bp

Table 2.6. Genomic regions covered by CGH array follow up of rare CNVs.

2.3.8 Statistical Analysis

We used Pearson's Chi^2 test to evaluate the probability that the frequency of cases with different CNVs over different regions of the genome compared to the frequency seen in both screened controls and WTCCC2 controls was due to chance(Pearson, n.d.). For calculations involving sample sizes of less than 10 we used Fisher's exact test(Fisher, n.d.). In a further analysis of CNV burden we used the --mperm and --cnv-test-2sided functions provided in the software PLINK (v1.07)(Purcell et al., 2007), using 10,000 permutations of case/control status to

calculate empirical 1 and 2 sided p values for association between cases and control cohorts.

2.4 Results

2,723 cases of recurrent depression, 348 screened controls and 4,828 unscreened controls from the WTCCC2 cohort passed quality control and were used in our main analysis. We used two methods to analyse our data:

1. A binary analysis comparing the number of samples with rare CNVs vs. the number of samples without rare CNVs in our case and control cohorts.
2. A burden analysis comparing the overall burden of rare CNVs in our case and control cohorts.

2.4.1 A Comparison of Samples With and Without Rare CNVs

SUMMARY: An enrichment of samples with CNVs seen in cases when compared to WTCCC2 controls, and also screened controls, is driven predominantly by samples with deletion CNVs and by deletion CNVs covering genic and exonic regions.

Our initial analysis aimed to compare the frequency of samples observed with a rare CNV against the frequency of samples observed without a CNV. We therefore categorised each sample according to whether a rare CNV was present (>0) or not present (0). This allowed us to calculate odds ratios with confidence intervals to gauge the relative significance of each result. We further categorised our samples by the region of the genome a CNV occurred in, namely

1. The whole genome.
2. Genic regions.
3. Exonic regions.
4. Intronic regions.
5. Intergenic regions.

This was achieved using Perl and bash scripts to filter and annotate our call files with genic and exonic coordinates and names supplied by the RefSeq(Pruitt, Tatusova, Klimke, & Maglott, 2009) gene and exon coordinate and annotation tracks (build hg18) available from the UCSC genome browser(W. J. Kent et al., 2002). Scripts can be viewed in the appendix.

We compared our cases firstly to our screened control group, and then to the Wellcome Trust Controls. As deletion CNVs may logically be expected to be more deleterious than duplication CNVs we stratified our analysis into deletion and duplication events after analysing all CNVs.

2.4.1.1 Analysis Across the Genome

Overall, the proportion of samples with a rare CNV was increased in cases when compared to screened controls and WTCCC2 controls ($p=0.019$, OR=1.31 (95% CI 1.04 - 1.64) and $p=0.025$, OR=1.11 (95% CI 1.01 - 1.23) respectively). After stratifying by type of CNV, deletions accounted for this change in both cohorts ($p=5.6x10^{-4}$, OR=1.52 (95% CI 1.20 - 1.93) and $p=7.7x10^{-6}$, OR =1.25 (95% CI 1.13 - 1.37) respectively). Tables 2.7 and 2.8 illustrate the frequency of samples with a CNV, stratified into deletions and duplications, in cases vs. screened

controls and cases vs. WTCCC2 controls respectively. Fig. 2.14 illustrates the proportion of samples with a CNV, stratified by type of CNV, across cohorts.

	Cases (N=2723)	**Screened Controls (N=348)**	**P value (Pearson's chi²)**	**Odds Ratio (95% CI)**
No. samples with a deletion (whole genome)	1123 41.2%	110 31.6%	$5.57x10^{-4}$	1.52 1.20 - 1.93
No. samples with a duplication (whole genome)	1047 38.5%	132 37.9%	0.85	1.02 0.81 - 1.28
No. samples with a deletion or duplication (whole genome)	1725 63.3%	198 56.9%	0.019	1.31 1.04 - 1.64

Table 2.7. Frequency of samples with deletion, duplication and any CNV in cases and screened controls with Pearson's chi² statistic and odds ratios with 95% confidence intervals.

	Cases (N=2723)	**WTCCC2 Controls (N=4828)**	**P value (Chi²)**	**Odds Ratio (95% CI)**
No. samples with a deletion (whole genome)	1123 41.2%	1740 36.0%	$7.70x10^{-6}$	1.25 1.13 - 1.37
No. samples with a duplication (whole genome)	1047 38.5%	1823 37.8%	0.55	1.03 0.93 - 1.13
No. samples with a deletion or duplication (whole genome)	1725 63.3%	2932 60.7%	0.025	1.12 1.01 - 1.23

Table 2.8. Frequency of samples with deletion, duplication and any CNV in cases and Wellcome Trust controls with Pearson's chi² statistic and odds ratios with 95% confidence intervals.

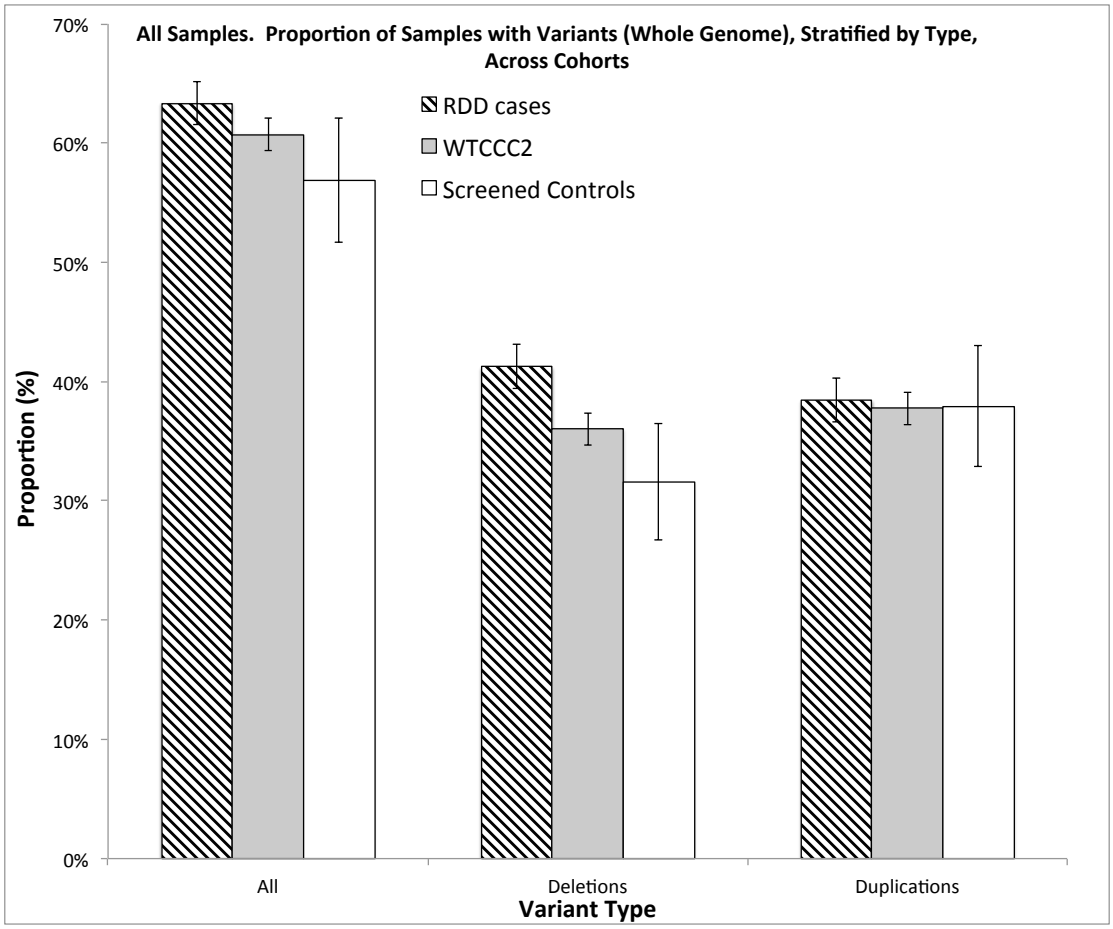


Fig. 2.14. Proportion of samples with variants across the whole genome, stratified by type, across cohorts.

2.4.1.2 Analysis Within Genic Regions

Within gene coding regions of the genome we found the frequency of samples with deletion CNVs was significantly increased in cases when compared to screened controls and WTCCC2 controls ($p=4.2x10^{-5}$, OR 1.78 (95% CI 1.35 - 2.35) and $p=7.2x10^{-6}$, OR 1.27 (95% CI 1.14 - 1.41) respectively). As with results for the whole genome, this effect was driven by an increased frequency of samples with deletions in both cohorts ($p = 4.24x10^{-5}$, OR 1.78 (95%CI 1.35 - 2.35) and $p= 7.17x10^{-6}$, OR 1.27 (95% CI 1.14 -1.41) for cases vs. screened controls and WTCCC2 controls respectively). Tables 2.9 and 2.10 illustrate the

frequency of samples with a genic CNV, stratified into deletions and duplications, in cases vs. screened controls and cases vs. WTCCC2 controls respectively. Fig. 2.15 illustrates the proportion of samples with a genic CNV, stratified by type of CNV, across cohorts.

	Cases (N=2723)	**Screened Controls (N=348)**	**P value (Pearson's chi²)**	**Odds Ratio (95% CI)**
No. samples with a genic deletion	811 29.8%	67 19.2%	4.24×10^{-5}	1.78 1.35 - 2.35
No. samples with a genic duplication	841 30.9%	105 30.1%	0.79	1.03 0.81 - 1.31
No. samples with a genic deletion or duplication	1392 51.1%	151 43.4%	0.0066	1.36 1.09 - 1.71

Table 2.9. Frequency of samples with deletion, duplication and any CNV in cases and screened controls with Pearson's chi² statistic and odds ratios with 95% confidence intervals.

	Cases (N=2723)	**WTCCC2 Controls (N=4828)**	**P value (Chi²)**	**Odds Ratio (95% CI)**
No. samples with a genic deletion	811 29.8%	1208 25.0%	7.17×10^{-6}	1.27 1.14 - 1.41
No. samples with a genic duplication	841 30.9%	1499 31.0%	0.88	0.99 0.90 - 1.10
No. samples with a genic deletion or duplication	1392 51.1%	2349 48.7%	0.040	1.10 1.00 - 1.21

Table 2.10. Frequency of samples with deletion, duplication and any CNV in cases and Wellcome Trust controls with Pearson's chi² statistic and odds ratios with 95% confidence intervals.

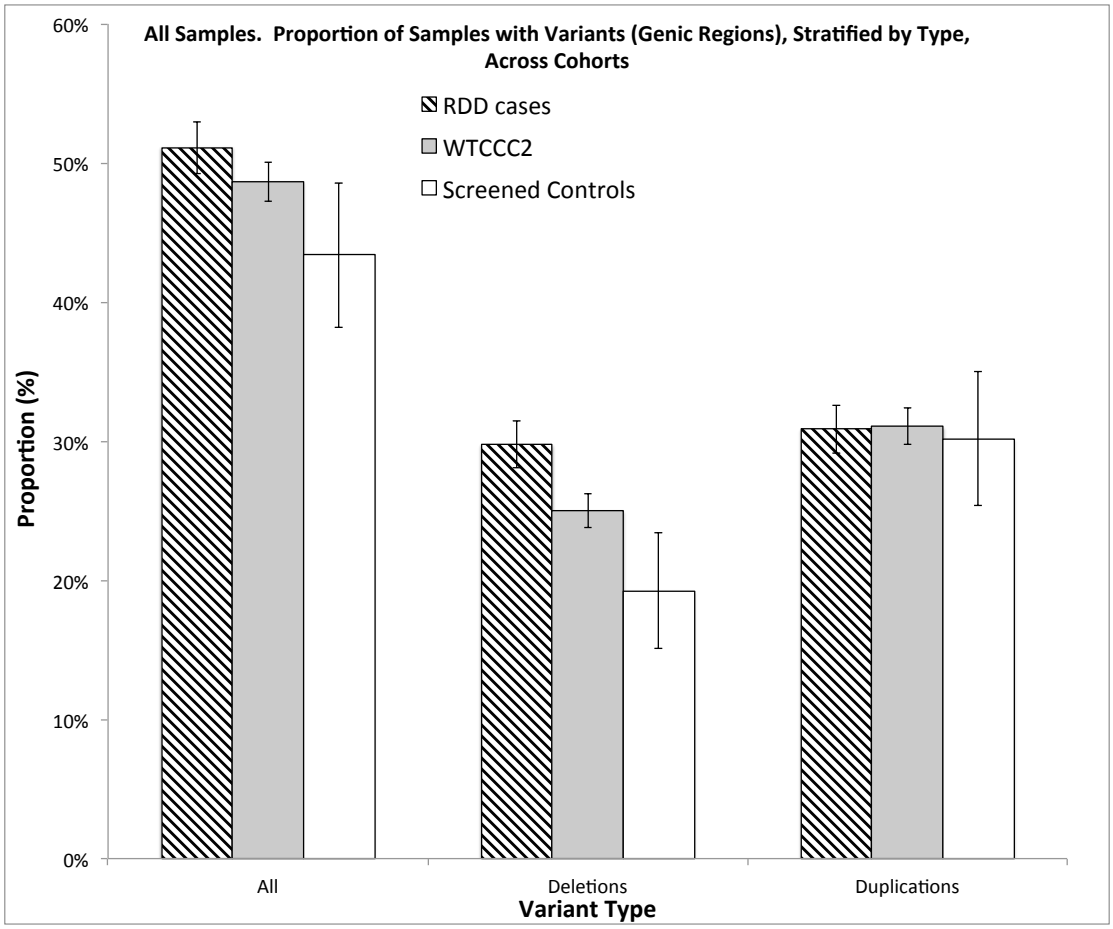


Fig 2.15. Proportion of samples with variants across genic regions of the genome, stratified by type, across cohorts.

2.4.1.3 Analysis Within Exonic Regions

We found an increased frequency of cases with an exonic deletion CNV when compared to screened controls and WTCCC2 controls ($p=1.70x10^{-5}$, OR 1.87 (95% CI 1.40 - 2.49) and $p=1.04x10^{-5}$, OR 1.27 (95% CI 1.14 - 1.41) respectively). The magnitude of the observed association suggests that samples with exonic deletions are partially driving our observed association between deletion CNVs and cases. Tables 2.11 and 2.12 illustrate the frequency of samples with an exonic CNV, stratified into deletions and duplications, in cases vs. screened controls and cases vs. WTCCC2 controls respectively. Fig. 2.16 illustrates the

proportion of samples with an exonic CNV, stratified by type of CNV, across cohorts.

	Cases (N=2723)	**Screened Controls (N=348)**	**P value (Pearson's chi²)**	**Odds Ratio (95% CI)**
No. samples with an exonic deletion*	774 28.4%	61 17.5%	$1.70x10^{-5}$	1.87 1.40 - 2.49
No. samples with an exonic duplication*	838 30.8%	105 30.1%	0.82	1.03 0.81 - 1.31
No. samples with an exonic deletion or duplication*	1362 50.0%	146 42.0%	0.0046	1.38 1.11 - 1.73

*Refers to samples with CNVs with any exonic coverage.

Table 2.11. Frequency of samples with deletion, duplication and any CNV in cases and screened controls with Pearson's chi² statistic and odds ratios with 95% confidence intervals.

	Cases (N=2723)	**WTCCC2 Controls (N=4828)**	**P value (Chi²)**	**Odds Ratio (95% CI)**
No. samples with an exonic deletion*	774 28.4%	1150 23.8%	$1.04x10^{-5}$	1.27 1.14 - 1.41
No. samples with an exonic duplication*	838 30.8%	1488 30.8%	0.97	1.00 0.90 - 1.10
No. samples with an exonic deletion or duplication*	1362 50.0%	2299 47.6%	0.045	1.27 1.14 - 1.41

*Refers to samples with CNVs with any exonic coverage.

Table 2.12. Frequency of samples with deletion, duplication and any CNV in cases and Wellcome Trust controls with Pearson's chi² statistic and odds ratios with 95% confidence intervals.

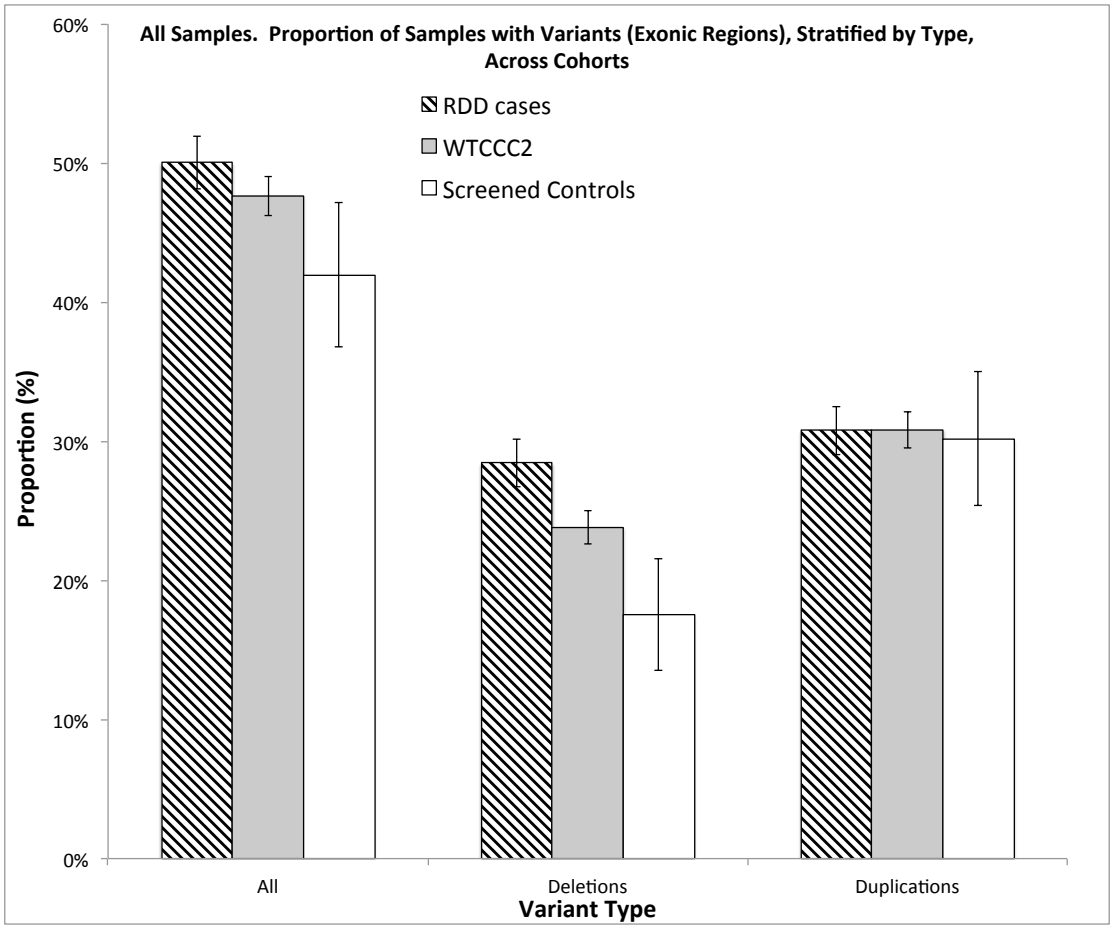


Fig 2.16. Proportion of samples with variants across exonic regions of the genome, stratified by type, across cohorts.

Analysis within intronic and intergenic regions did not yield significant results, and can be viewed in the appendix.

2.4.2 Analysis by Gender

We stratified our data by gender, both to analyse for gender-specific effects and also because gender proportions differ between our own samples and the WTCCC2 controls. Table 2.13 compares the frequency of samples with CNVs in cases and screened controls, stratified by CNV type and gender. Table 2.14 compares the frequency of samples with CNVs in cases and WTCCC2 controls, stratified by CNV type and gender. Whilst we see a generally similar picture to

previous results, with significantly more female cases than controls harbouring a deletion, there is no significant difference between the proportion of male samples who have a deletion CNV when cases are compared to screened controls.

	Cases N=2723		Screened Controls N=348		P value (Chi²)		Odds Ratio 95% CI	
	Males N=817	Females N=1906	Males N=115	Females N=233	Males	Females	Males	Females
No. samps with a del	334 40.9%	789 41.4%	41 35.7%	69 29.6%	0.28	$5.3x10^{-4}$	1.25 0.83-1.87	1.68 1.25-2.25
No. samps with a dup	319 39.0%	728 38.2%	42 36.5%	90 38.6%	0.60	0.90	1.11 0.74-1.67	0.98 0.74-1.30
No. samps with a del/dup	516 63.2%	1209 63.3%	64 55.7%	134 57.5%	0.12	0.078	1.37 0.92-2.02	1.28 0.97-1.69

Table 2.13. Comparison of the frequency of samples with CNVs in cases and screened controls, stratified by CNV type, and gender.

	Cases N=2723		WTCCC2 Controls N=4828		P value (Chi²)		Odds Ratio 95% CI	
	Males N=817	Females N=1906	Males N=2412	Females N=2416	Males	Females	Males	Females
No. samps with a del	334 40.9%	789 41.4%	877 36.3%	863 35.7%	0.021	$1.38x10^{-4}$	1.21 1.03-1.42	1.27 1.12-1.44
No. samps with a dup	319 39.0%	728 38.2%	904 37.5%	919 38.0%	0.43	0.92	1.07 0.91-1.26	1.01 0.89-1.14
No. samps with a del/dup	516 63.2%	1209 63.3%	1461 60.6%	1471 60.9%	0.19	0.087	1.12 0.95-1.31	1.11 0.98-1.26

Table 2.14. Comparison of the frequency of samples with CNVs in cases and WTCCC2 controls, stratified by CNV type, and gender.

2.4.3 High QC Analyses

SUMMARY: We re-analysed our data after removing the worst performing 10% of samples by LRRSD and BAFSD across cohorts. Our analyses remain significant, but with significantly reduced effect, when these samples are removed.

Accurate comparisons of CNV frequency between groups of samples can be confounded by differences in the numbers of false positive calls made in the worst performing samples. To try and account for this we re-analysed our data post hoc after removing samples falling above the 90th centile of the LRRSD and the BAFSD for the entire sample. This restricted analysis to a sample set with a LRRSD of less than 0.2241 and a BAFSD of less than 0.039 (Table 2.15). More samples from the screened control group fail this higher QC threshold than cases or population controls.

Cohort	QC Statistic	No. Samples (% of original)	Mean	SD	Min	Max
Cases	**BAFSD**	2207 (81.0%)	0.032	0.0029	0.025	0.039
	LRRSD		0.164	0.0264	0.111	0.224
Screened Controls	**BAFSD**	257 (73.9%)	0.033	0.0026	0.026	0.039
	LRRSD		0.156	0.0246	0.111	0.223
WTCCC2 Controls	**BAFSD**	4212 (87.2%)	0.031	0.0040	0.021	0.039
	LRRSD		0.156	0.0273	0.090	0.224

Table 2.15. Sample QC characteristics after removal of the 90-100th percentile of samples as defined by LRRSD and BAFSD across all cohorts.

A full set of results comparing cases to screened controls and WTCCC2 controls is given in tables 2.16 and 2.17 respectively. In summary, across the whole genome (Fig. 2.17) deletions remain significant when cases are compared to screened controls and WTCCC2 controls but the magnitude of association is reduced by a significant degree (p=0.017, OR 1.41 (95% CI 1.06 - 1.87) and

$p=5.5x10^{-3}$, OR 1.16 (95% CI 1.05 - 1.30) respectively). This reflects partly a reduction in power and partly the removal of samples more likely to contribute false positive calls.

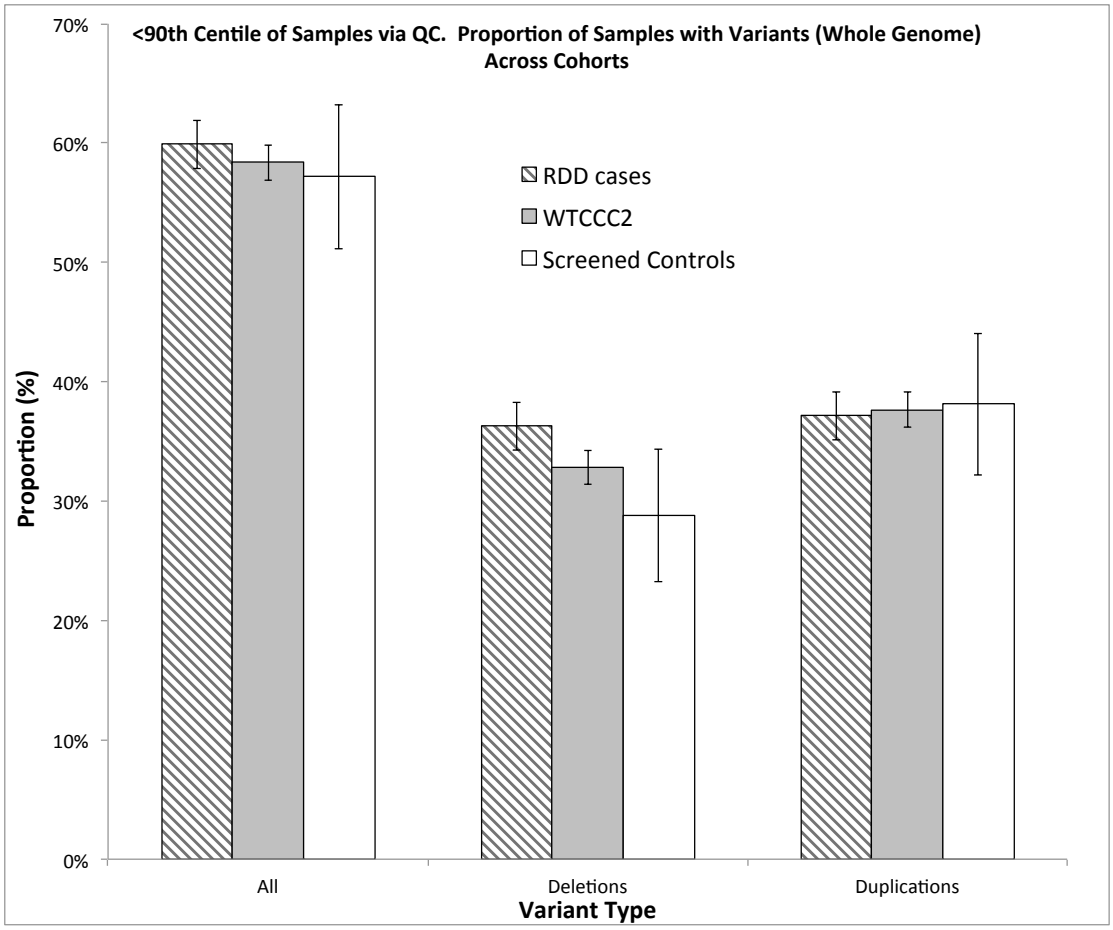


Fig 2.17. Proportion of high QC (LRRSD < 0.2241 & BAFSD < 0.0390) samples with deletion and duplication CNVs in all regions of the genome in screened controls, WTCCC2 controls and cases. Error bars represent 95% confidence intervals.

Across exonic regions (Fig. 2.18) deletions also remain significant when cases are compared to screened controls and WTCCC2 controls (p=0.0060, OR 1.63 (95% CI 1.15 - 2.31) and p=0.043, OR 1.14 (95% CI 1.00 - 1.29).

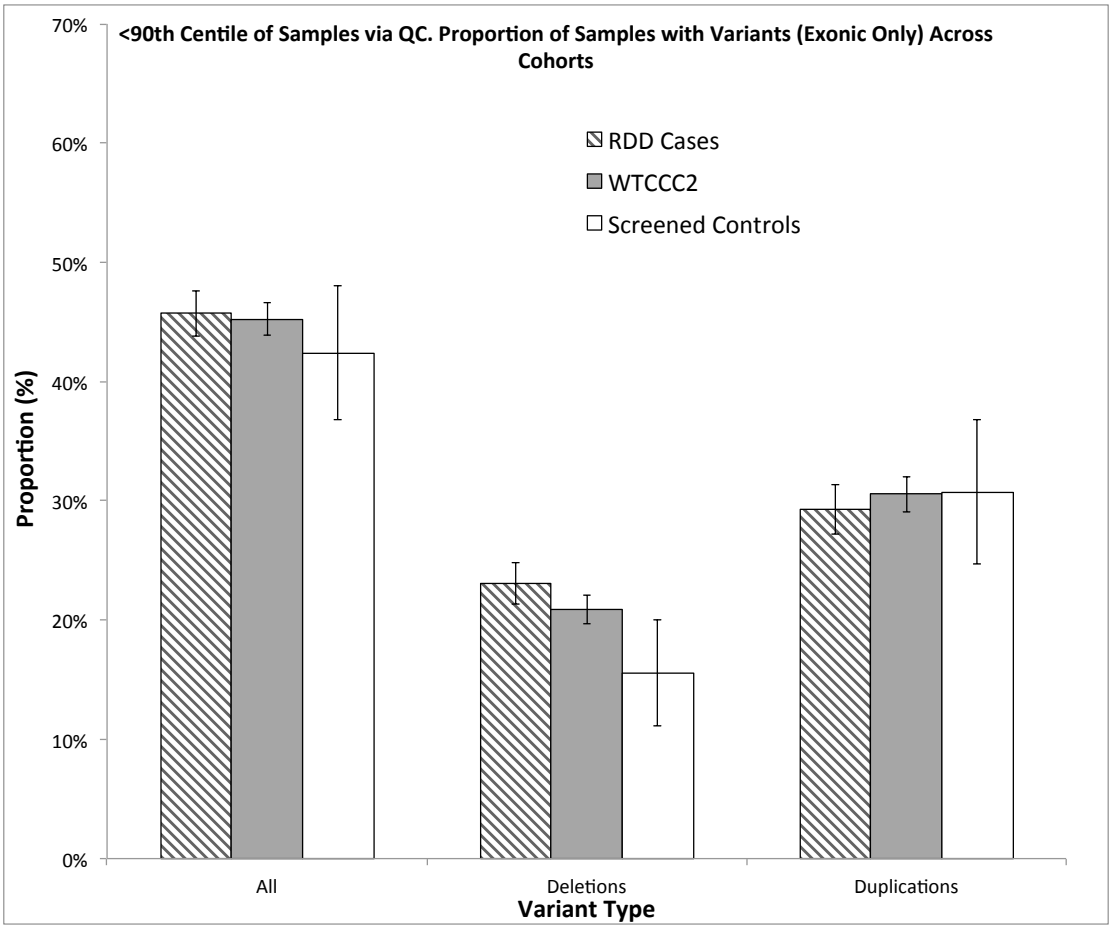


Fig 2.18. Proportion of high QC (LRRSD < 0.2241 & BAFSD < 0.039) samples with deletion and duplication CNVs in exonic regions of the genome in screened control , WTCCC2 controls and cases. Error bars represent 95% confidence intervals.

	Cases N=2207	Screened Controls N=257	P value (P'son's chi²)	Odds Ratio (95% CI)
No. samples with a deletion (whole genome)	801 (36.3%)	74 (28.8%)	0.017	1.41 (1.06 - 1.87)
No. samples with a duplication (whole genome)	820 (37.1%)	98 (38.1%)	0.76	0.95 (0.74 - 1.25)
No. samples with a deletion or duplication (whole genome)	1322 (59.9%)	147 (57.2%)	0.40	1.18 (0.86 - 1.45)
No. samples with an intergenic deletion	359 (16.3%)	40 (15.6%)	0.77	1.05 (0.74 - 1.50)
No. samples with an intergenic duplication	272 (12.3%)	32 (12.5%)	0.95	0.99 (0.67 - 1.45)
No. samples with an intergenic deletion or duplication	587 (26.6%)	67 (26.1%)	0.86	1.03 (0.77 - 1.38)
No. samples with a genic deletion	539 (24.4%)	41 (16.0%)	0.0025	1.70 (1.20 - 2.41)
No. samples with a genic duplication	649 (29.4%)	79 (30.7%)	0.66	0.94 (0.71 - 1.24)
No. samples with a genic deletion or duplication	1033 (46.8%)	110 (42.8%)	0.22	1.18 (0.91 - 1.53)
No. samples with an intronic deletion**	40 (1.8%)	3 (1.2%)	0.62***	1.56 (0.51 - 4.79)
No. samples with an intronic duplication**	19 (0.9%)	2 (0.8%)	0.68***	1.38 (0.57 - 3.38)
No. samples with an intronic deletion or duplication**	59 (2.6%)	5 (1.9%)	1.00***	1.10 (0.28 - NaN)
No. samples with an exonic deletion*	510 (23.1%)	40 (15.6%)	0.0060	1.63 (1.15 - 2.31)
No. samples with an exonic duplication*	646 (29.2%)	79 (30.7%)	0.62	0.93 (0.70 - 1.23)
No. samples with an exonic deletion or duplication*	1009 (45.7%)	109 (42.4%)	0.31	1.14 (0.88 - 1.48)

*Refers to samples with CNVs with any exonic coverage. **Refers to samples with CNVs with only intronic coverage. ***Statistics including samples with N≤10 are calculated with Fisher's exact method rather than Pearson's chi²

Table 2.16. Frequency of high QC samples with CNVs, stratified by type, in cases compared to screened controls in non-gene coding, gene coding, intronic, exonic and all regions of the genome

	Cases N=2207	WTCCC2 Controls N=4212	P value (P'son's chi2)	Odds Ratio (95% CI)
No. samples with a deletion (whole genome)	801 (36.3%)	1383 (32.8%)	0.0055	1.16 (1.05 - 1.30)
No. samples with a duplication (whole genome)	820 (37.1%)	1586 (37.7%)	0.69	0.98 (0.88 - 1.09)
No. samples with a deletion or duplication (whole genome)	1322 (59.9%)	2459 (58.4%)	0.24	1.06 (0.96 - 1.18)
No. samples with an intergenic deletion	359 (16.3%)	626 (14.9%)	0.14	1.11 (0.97 - 1.28)
No. samples with an intergenic duplication	272 (12.3%)	429 (10.2%)	0.0091	1.23 (1.05 - 1.46)
No. samples with an intergenic deletion or duplication	587 (16.6%)	1000 (23.7%)	0.018	1.16 (1.03 - 1.31)
No. samples with a genic deletion	539 (24.4%)	932 (22.1%)	0.038	1.14 (1.01 - 1.28)
No. samples with a genic duplication	649 (29.4%)	1296 (30.8%)	0.26	0.94 (0.84 - 1.05)
No. samples with a genic deletion or duplication	1033 (46.8%)	1949 (46.3%)	0.68	1.02 (0.92 - 1.13)
No. samples with an intronic deletion**	40 (1.8%)	80 (1.9%)	0.81	0.95 (0.65 - 1.40)
No. samples with an intronic duplication**	19 (0.9%)	34 (0.8%)	0.82	1.07 (0.61 - 1.86)
No. samples with an intronic deletion or duplication**	59 (2.6%)	114 (2.7%)	0.94	0.99 (0.72 - 1.36)
No. samples with an exonic deletion*	510 (23.1%)	881 (20.9%)	0.043	1.14 (1.00 - 1.29)
No. samples with an exonic duplication*	646 (29.2%)	1287 (30.6%)	0.29	0.94 (0.84 - 1.05)
No. samples with an exonic deletion or duplication*	1009 (45.7%)	1906 (45.3%)	0.72	1.02 (0.92 - 1.13)

*Refers to samples with CNVs with any exonic coverage
**Refers to samples with CNVs with only intronic coverage
Table 2.17. Frequency of high QC samples with CNVs, stratified by type, in cases compared to WTCCC2 controls in non-gene coding, gene coding, intronic, exonic and all regions of the genome.

2.4.4 UK Population Analyses

SUMMARY: An analysis restricted to samples with exclusively UK origin found a similar effect to our main analysis, suggesting that our results are not biased by population-attributable differences in rare CNV frequency across cohorts.

Differences in frequencies of rare CNVs may occur between different populations, depending on the degree of purifying selection selecting against deleterious CNVs(F. Zhang, Gu, Hurles, & Lupski, 2009b). Whilst our samples are all of European heritage, approximately half of our case cohort is of non-UK origin, whilst both our control cohorts are of almost exclusively UK origin. To check whether this affected our association we re-analysed our data after restricting our samples to those individuals with exclusively UK origin. A full set of results comparing cases to screened controls and WTCCC2 controls is given in tables 2.18 and 2.19 respectively. In summary, deletion CNVs across the whole genome remain significant when cases are compared to screened controls and WTCCC2 controls ($p=0.0045$, OR 1.44 (95% CI 1.12 - 1.84) and $p=0.0044$, OR 1.19 (95% CI 1.06 - 1.35) respectively) (Fig. 2.19). Exonic deletion CNVs also remain significant ($p=6.97 \times 10^{-5}$, OR 1.83 (95% CI 1.36 - 2.48) and $p=0.0012$, OR 1.25 (95% CI 1.09 - 1.42) respectively (Fig. 2.20).

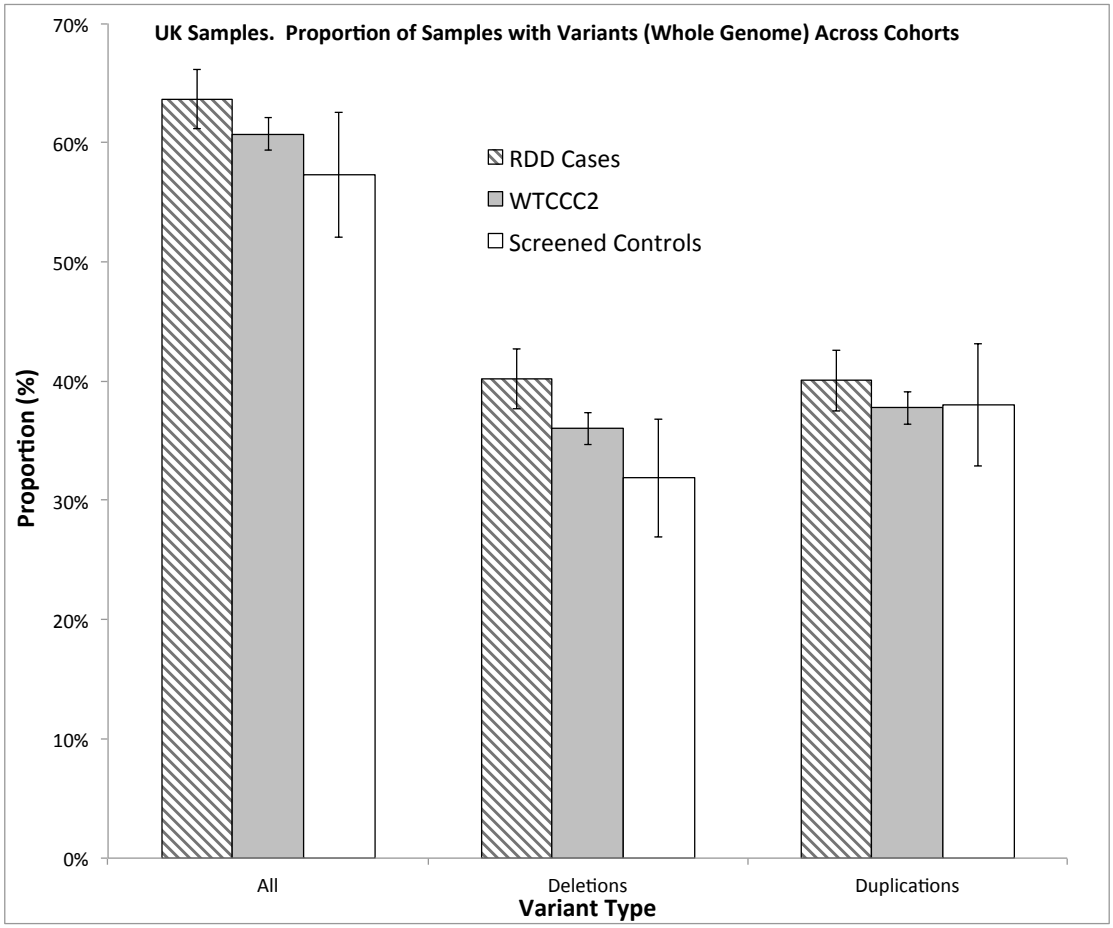


Fig 2.19. Proportion of UK-only samples with deletion and duplication CNVs in all regions of the genome in screened controls, WTCCC2 controls and cases. Error bars represent 95% confidence intervals.

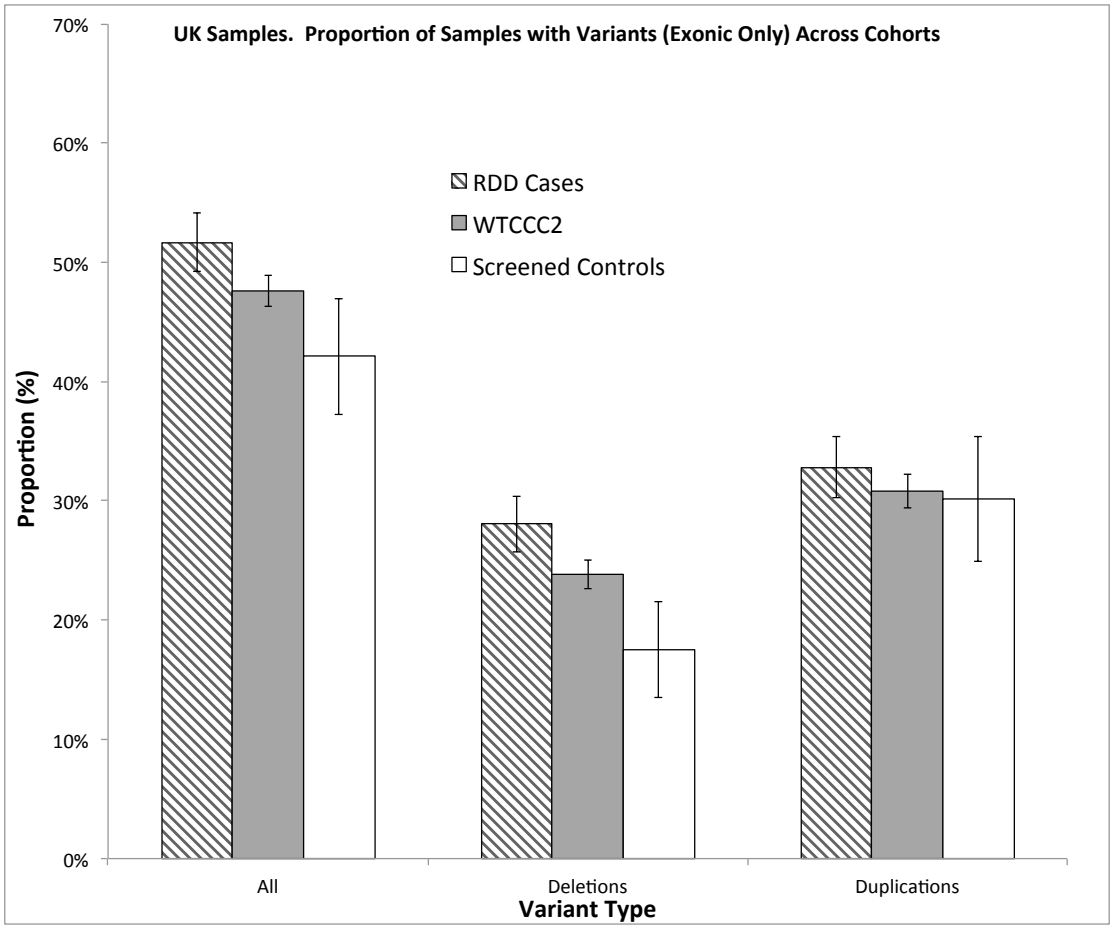


Fig 2.20. Proportion of UK-only samples with deletion and duplication CNVs in exonic regions of the genome in screened controls , WTCCC2 controls and cases. Error bars represent 95% confidence intervals.

	Cases N=1426	Screened Controls N=342	P value (P'son's chi^2)	Odds Ratio (95% CI)
No. samples with a deletion (whole genome)	573 (40.2%)	109 (31.9%)	0.0046	1.44 (1.12 - 1.84)
No. samples with a duplication (whole genome)	571 (40.0%)	130 (38.0%)	0.49	1.09 (0.85 - 1.39)
No. samples with a deletion or duplication (whole genome)	908 (63.7%)	196 (57.3%)	0.029	1.31 (1.03 - 1.66)
No. samples with an intergenic deletion	270 (18.9%)	59 (17.3%)	0.47	1.12 (0.82 - 1.53)
No. samples with an intergenic duplication	173 (12.1%)	42 (12.3%)	0.94	0.99 (0.69 - 1.41)
No. samples with an intergenic deletion or duplication	406 (28.5%)	93 (27.2%)	0.64	1.07 (0.82 - 1.39)
No. samples with a genic deletion	421 (29.5%)	66 (19.3%)	0.00014	1.75 (1.31 - 2.34)
No. samples with a genic duplication	470 (33.0%)	103 (30.1%)	0.31	1.14 (0.88 - 1.47)
No. samples with a genic deletion or duplication	754 (52.9%)	149 (43.6%)	0.0012	1.45 (1.15 - 1.84)
No. samples with an intronic deletion**	32 (2.2%)	8 (2.3%)	0.84***	0.96 (0.45 - 2.06)
No. samples with an intronic duplication**	19 (1.3%)	2 (0.6%)	0.40***	2.30 (0.59 - NaN)
No. samples with an intronic deletion or duplication**	51 (3.6%)	10 (2.9%)	0.55	1.23 (0.63 - 2.42)
No. samples with an exonic deletion*	400 (28.1%)	60 (17.5%)	$6.97x10^{-5}$	1.83 (1.36 - 2.48)
No. samples with an exonic duplication*	468 (32.8%)	103 (30.1%)	0.34	1.13 (0.88 - 1.46)
No. samples with an exonic deletion or duplication*	737 (51.7%)	144 (42.1%)	0.0015	1.47 (1.16 - 1.87)

*Refers to samples with CNVs with any exonic coverage. **Refers to samples with CNVs with only intronic coverage. ***Statistics including samples with N≤10 are calculated with Fisher's exact method rather than Pearson's chi^2.
Table 2.18. Frequency of UK-only samples with variants, stratified by CNV type, in cases compared to screened controls in non-gene coding, gene coding, intronic, exonic and all regions of the genome.

	Cases N=1426	WTCCC2 Controls N=4828	P value (P'son's chi^2)	Odds Ratio (95% CI)
No. samples with a deletion (whole genome)	573 (40.2%)	1740 (36.0%)	0.0044	1.19 (1.06 - 1.35)
No. samples with a duplication (whole genome)	571 (40.0%)	1823 (37.8%)	0.12	1.10 (0.98 - 1.24)
No. samples with a deletion or duplication (whole genome)	908 (63.7%)	2932 (60.7%)	0.045	1.13 (1.00 - 1.28)
No. samples with an intergenic deletion	270 (18.9%)	846 (17.5%)	0.22	1.10 (0.94 - 1.28)
No. samples with an intergenic duplication	173 (12.1%)	488 (10.1%)	0.029	1.23 (1.02 - 1.48)
No. samples with an intergenic deletion or duplication	406 (28.5%)	1260 (26.1%)	0.075	1.13 (0.99 - 1.29)
No. samples with a genic deletion	421 (29.5%)	1208 (25.0%)	$6.65x10^{-4}$	1.26 (1.10 - 1.43)
No. samples with a genic duplication	470 (33.0%)	1499 (31.0%)	0.17	1.09 (0.96 - 1.24)
No. samples with a genic deletion or duplication	754 (52.9%)	2349 (48.7%)	0.0051	1.18 (1.05 - 1.33)
No. samples with an intronic deletion**	32 (2.2%)	1082 (22.4%)	0.67	0.92 (0.62 - 1.37)
No. samples with an intronic duplication**	19 (1.3%)	1437 (29.8%)	0.084	1.62 (0.94 - 2.78)
No. samples with an intronic deletion or duplication**	51 (3.6%)	2212 (45.8%)	0.52	1.11 (0.81 - 1.53)
No. samples with an exonic deletion*	400 (28.1%)	1150 (23.8%)	0.0011	1.25 (1.09 - 1.42)
No. samples with an exonic duplication*	468 (32.8%)	1488 (30.8%)	0.15	1.10 (0.97 - 1.24)
No. samples with an exonic deletion or duplication*	737 (51.7%)	2299 (47.6%)	0.0070	1.18 (1.05 - 1.32)

*Refers to samples with CNVs with any exonic coverage
**Refers to samples with CNVs with only intronic coverage
Table 2.19. Frequency of UK-only samples with variants, stratified by variant type, in cases compared to WTCCC2 controls in non-gene coding, gene coding, intronic, exonic and all regions of the genome.

2.4.5 Regions Previously Associated with Schizophrenia

We undertook a further, more detailed review of other CNVs that have been associated with schizophrenia, and were recently covered in a review by Kirov(Kirov, 2010). Deletions or duplications (but usually not deletions and duplications in the same area) in specific areas have been documented as being associated with schizophrenia, as well as other disorders such as autism, sometimes with the reciprocal CNV being associated with schizophrenia and autism respectively(Shane E McCarthy et al., 2009). Schizophrenia and recurrent depression may or may not share genetic similarities, but given the observation of pleiotropy in CNV association, we present here counts of CNVs, both duplications and deletions, regardless of prior evidence of association with a specific CNV type. CNV boundaries also vary according to array type and detection methodology used, as well as sample, therefore we would not expect precise overlap, and in any event CNV call sizes and estimated breakpoints are frequently slightly dissimilar between individuals when called on SNP microarrays. Thus we present counts of CNVs covering ≥90% of the published region, <90% of the published region, and totals with proportions, stratified by CNV type for each region in table 2.20.

We found that our cases were significantly enriched with CNVs in regions previously implicated in schizophrenia when compared to screened controls and when both deletions and duplications were considered together (p=0.0190, OR=3.21 (95%CI 1.07 - 9.68)). However no significant difference was observed when cases were compared to the WTCCC2 controls (p=0.88, OR=0.98 (95%CI 0.73-1.30)). No individual region was significantly associated.

Locus & Position (Mb)	**Type**	**CNV Region Cov'ge**	**Number of Samples with CNVs**			**Freq. from Kirov**(Kirov, 2010)
			Cases N=2723	**WTCCC2 Controls N=4828**	**Screened Controls N=348**	**% Cases | % Controls**
1q21.1 144.9-146.3	Del	≥90%	2 (0.07%)	0 (0%)	0 (0%)	0.2 | 0.02
		<90%	0 (0%)	2 (0.04%)	0 (0%)	
		Total	2 (0.07%)	2 (0.04%)	0 (0%)	
	Dup	≥90%	0 (0%)	1 (0.02%)	0 (0%)	NS
		<90%	1 (0.04%)	0 (0%)	0 (0%)	
		Total	1 (0.04%)	1 (0.02%)	0 (0%)	
2p16.3 50.0-51.2	Del	≥90%	0 (0%)	0 (0%)	0 (0%)	0.2 | 0.04
		<90%	2 (0.07%)	4 (0.08%)	0 (0%)	
		Total	2 (0.07%)	4 (0.08%)	0 (0%)	
	Dup	≥90%	0 (0%)	0 (0%)	0 (0%)	NS
		<90%	0 (0%)	0 (0%)	0 (0%)	
		Total	0 (0%)	0 (0%)	0 (0%)	
15q11.2 20.2-20.8	Del	≥90%	9 (0.33%)	20 (0.41%)	1 (0.29%)	0.6 | 0.22
		<90%	1 (0.04%)	1 (0.02%)	0 (0%)	
		Total	10 (0.37%)	21 (0.43%)	1 (0.29%)	
	Dup	≥90%	10 (0.37%)	14 (0.29%)	1 (0.29%)	NS
		<90%	3 (0.11%)	3 (0.06%)	0 (0%)	
		Total	13 (0.48%)	17 (0.35%)	1 (0.29%)	
15q13.3 29-30.3	Del	≥90%	0 (0%)	0 (0%)	0 (0%)	0.2 | 0.02
		<90%	0 (0%)	0 (0%)	0 (0%)	
		Total	0 (0%)	0 (0%)	0 (0%)	
	Dup	≥90%	0 (0%)	2 (0.04%)	0 (0%)	NS
		<90%	18 (0.66%)	44 (0.91%)	1 (0.29%)	
		Total	18 (0.66%)	46 (0.95%)	1 (0.29%)	
16p13.1 15-16.4	Del	≥90%	3 (0.11%)	3 (0.06%)	0 (0%)	NS
		<90%	1 (0.04%)	2 (0.04%)	0 (0%)	
		Total	4 (0.15%)	5 (0.1%)	0 (0%)	
	Dup	≥90%	1 (0.04%)	3 (0.06%)	0 (0%)	0.3-0.5 | 0.1-0.25
		<90%	6 (0.22%)	7 (0.14%)	0 (0%)	
		Total	7 (0.26%)	10 (0.21%)	0 (0%)	
16p11.2 29.5-30.1	Del	≥90%	2 (0.07%)	3 (0.06%)	0 (0%)	NS
		<90%	0 (0%)	1 (0.02%)	0 (0%)	
		Total	2 (0.07%)	4 (0.08%)	0 (0%)	
	Dup	≥90%	3 (0.11%)	1 (0.02%)	0 (0%)	0.3 | 0.03
		<90%	0 (0%)	1 (0.02%)	0 (0%)	
		Total	3 (0.11%)	2 (0.04%)	0 (0%)	
22q11.2 17.4-19.8	Del	≥90%	1 (0.04%)	0 (0%)	0 (0%)	0.2 | NS
		<90%	2 (0.07%)	3 (0.06%)	0 (0%)	
		Total	3 (0.11%)	3 (0.06%)	0 (0%)	
	Dup	≥90%	0 (0%)	2 (0.04%)	0 (0%)	NS
		<90%	9 (0.33%)	17 (0.35%)	0 (0%)	
		Total	9 (0.33%)	19 (0.39%)	0 (0%)	
All Regions	Del	Total	24 (0.88%)	37 (0.77%)	1 (0.29%)	NS
	Dup	Total	50 (1.84%)	97 (2.01%)	2 (0.57%)	
	Del/ Dup	Total	74 (2.72%)	134 (2.78%)	3 (0.86%)	

Table 2.20. Numbers and percentages of samples with CNVs in genomic regions previously implicated in schizophrenia. NS= not stated.

2.4.6 Analysis of Genomic Regions 1q21.1, 15q13.3 and 22q11.2

There is strong evidence for association between psychiatric disorders and variants seen at the genomic regions 1q21.1, 15q13.3 and 22q11.2(International Schizophrenia Consortium, 2008; Kirov, Grozeva, Norton, Ivanov, Mantripragada, Holmans, Craddock, Owen, et al., 2009a; Shinawi et al., 2009). We compared the frequencies of samples with CNVs seen in any part of these areas between our case and control cohorts, calculating Pearson's chi^2 or Fisher's exact tests as appropriate to probe for statistical significance. We present screenshots of the UCSC browser for each region, and selected plots of LRR and BAF values for selected calls.

2.4.6.1 1q21.1

Within the 1q21.1 region (chr1:142,400,001-148,000,000) there were no statistically significant associations when cases were compared to either of the control groups. Two cases (neither with psychotic symptoms) have an 900KB deletion CNV (Fig. 2.22) between two areas of segmental duplication. In both cases the deletion probably extends beyond the area of segmental duplication around 146.0Mb (see far right of each plot) but this has been called as a separate event by PennCNV and subsequently excluded from our analysis.

Cohort/Variant Type		All CNVs	Deletions	Duplications
WTCCC2 Controls	**1958 Birth Cohort**	12 0.48%	5 0.20%	7 0.28%
	National Blood Service	8 0.34%	4 0.17%	4 0.17%
Cases		15 0.55%	3 0.11%	12 0.44%
Screened Controls		1 0.29%	0 0%	1 0.29%
WTCCC2 Controls		20 0.41%	9 0.19%	11 0.22%

Table 2.21. Frequency of samples with CNVs in the 1q21.1 region (chr1:142,400,001-148,000,000) hg18 build.

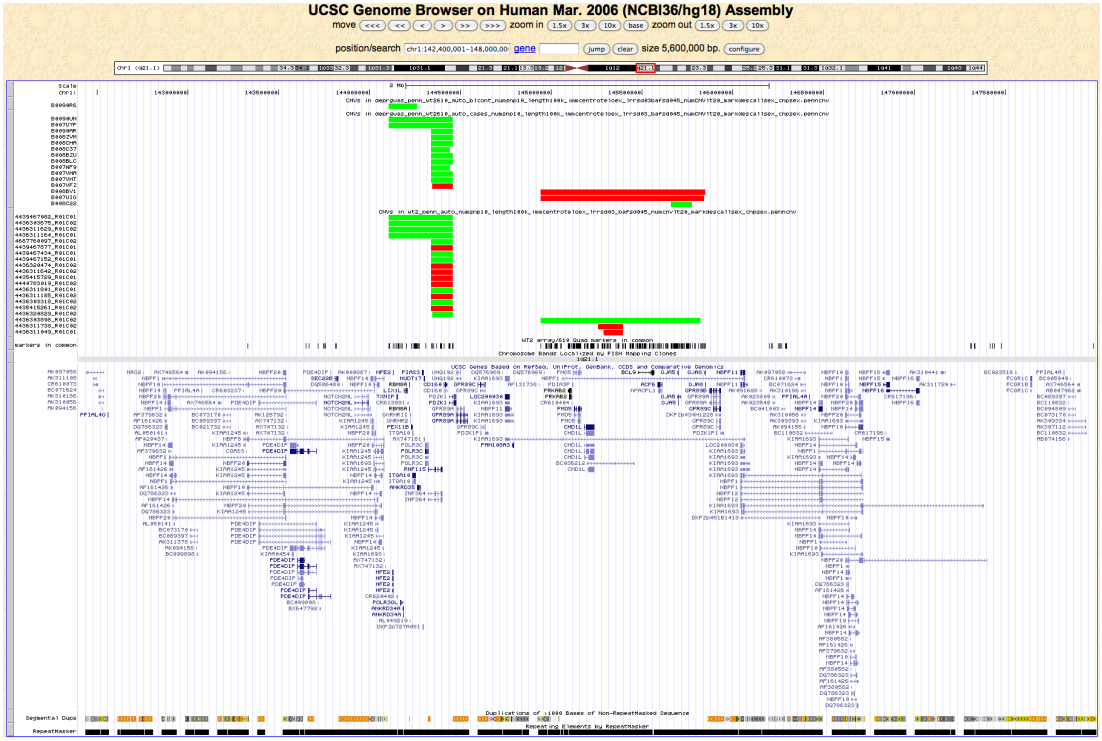


Fig 2.21. UCSC Genome Browser. Frequency of samples with deletion (red lines) CNVs and duplication (green lines) CNVs in screened controls (upper tier), cases (middle tier) and WTCCC2 controls (lower tier) in Chr1q21.1. Two cases have a 900KB deletion CNV (Fig 2.22).

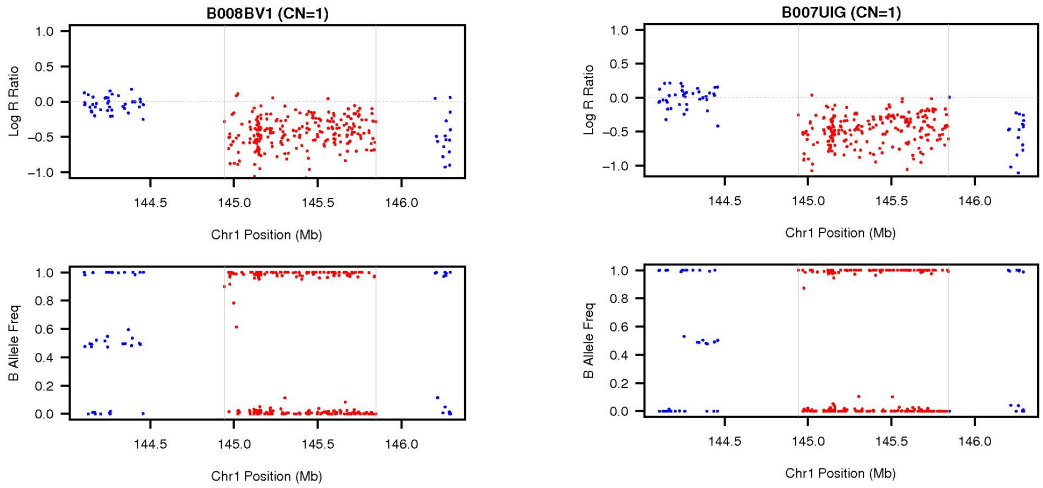


Fig 2.22. Two cases have large 900KB deletion CNVs surrounding areas of segmental duplication in 1q21.1.

2.4.6.2 15q13.3

Within the 15q13.3 region (chr15:29,000,001-31,400,000) a Pearson's Chi^2 test for the frequency of all samples with 15q13.3 CNVs between cases and NBS controls yielded a p value of 0.044, with the NBS sample having more variants than the sample. No other comparisons were statistically significant.

Cohort/Variant Type		All CNVs	Deletions	Duplications
WTCCC2 Controls	**1958 Birth Cohort**	22 0.89%	0 0%	22 0.89%
	National Blood Service	27 1.2%	2 0.08%	25 1.1%
Cases		17 0.62%	0 0%	17 0.62%
Screened Controls		1 0.29%	0 0%	1 0.29%
WTCCC2 Controls		49 1.0%	2 0.04%	47 0.97%

Table 2.22. Frequency of samples with CNVs in the 15q13.3 region (chr15:29,000,001-31,400,000).

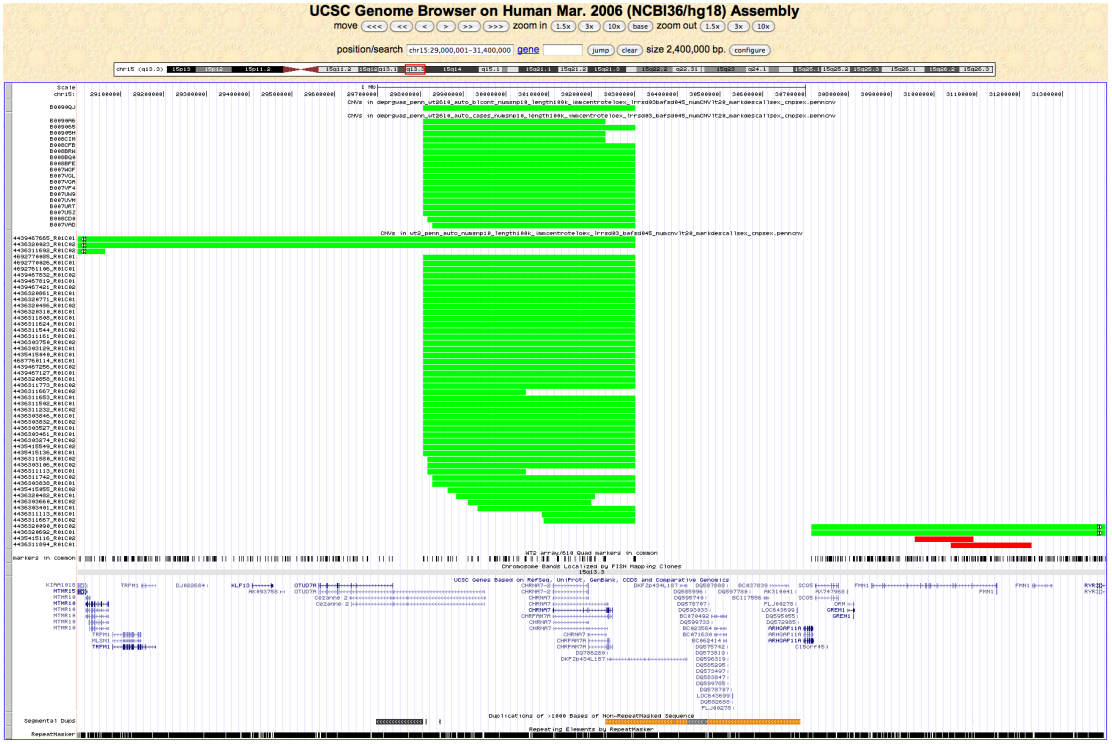


Fig 2.23. UCSC Genome Browser. Frequency of samples with deletion (red lines) CNVs and duplication (green lines) CNVs in screened controls (upper tier), cases (middle tier) and WTCCC2 controls (lower tier) in 15q13.3.

2.4.6.3 22q11.2

We performed a 1 sided Fisher's exact test for the frequency of samples with deletion CNVs of 22q11.2 (chr22:16,300,001-24,300,000), as prior research had implicated this region in both psychiatric disorders(Karayiorgou et al., 2010), including mood disturbance(Jolin et al., 2009; Papolos et al., 1996), and the velocardiofacial syndrome(Robin & Shprintzen, 2005), which is strongly associated with development of schizophrenia. When the frequency of samples with deletion CNVs in cases was compared to national blood service controls a p value of 0.045 was obtained. Odds ratios cannot be calculated as there were no deletion CNVs seen in the screened control cohort. No other comparisons were statistically significant.

Cohort/Variant Type		All CNVs	Deletions	Duplications
WTCCC2 Controls	**1958 Birth Cohort**	9 0.36%	4 0.16%	5 0.20%
	National Blood Service	11 0.47%	0 0%	11 0.47%
Cases		18 0.66%	5 0.18%	13 0.48%
Screened Controls		0 0%	0 0%	0 0%

Table 2.23. Frequency of variants in the 22q11.2 region (chr22:16,300,001-24,300,000).

Inspection of CNV calls in the UCSC genome browser (Fig. 2.24) reveals 1 individual without psychotic symptoms with a large deletion CNV of 22q11.2, and 4 individuals none of whom demonstrated psychotic symptoms with smaller deletion CNVs of 22q11.2.

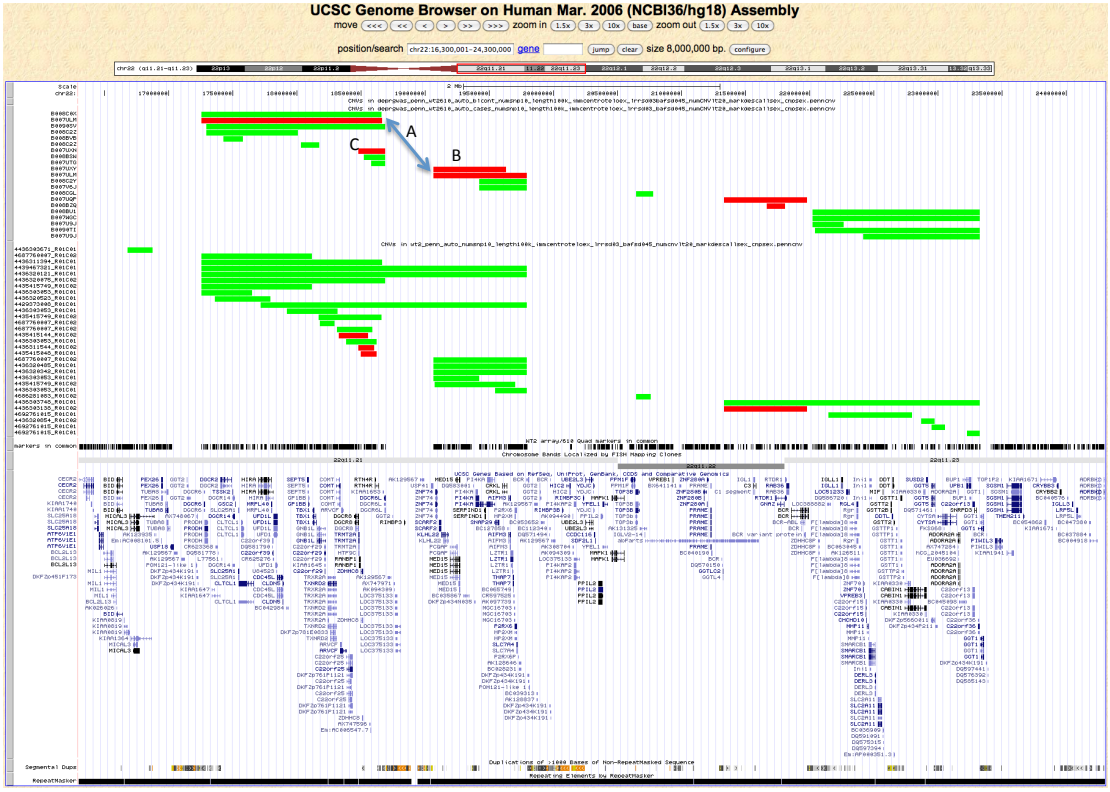


Fig 2.24. UCSC Genome Browser. Frequency of samples with deletion (red lines) CNVs and duplication (green lines) CNVs in screened controls (upper tier- note no CNVs are seen), cases (middle tier) and WTCCC2 controls (lower tier) in Chr22q11.2. Note that there are no deletions or duplications in the screened control group. Variant A is split around an area of segmental duplication (Fig. 2.25).

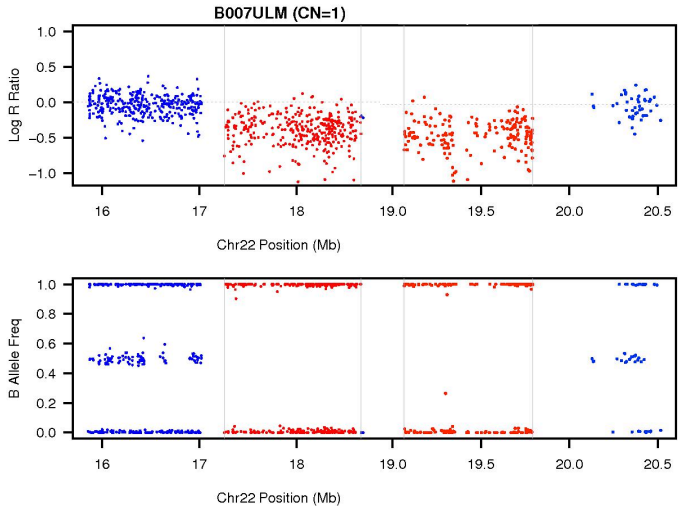


Fig. 2.25. Variant A in Fig 2.24 is a large deletion CNV in chromosome 22q11.2 called around an area of segmental duplication.

2.4.7 CNV Burden Analysis

SUMMARY: The average CNV size remained consistent across cohorts. Screened control samples have a particularly low burden of deletion CNVs, however in this analysis method WTCCC2 controls and cases have a broadly similar deletion burden, although the proportion of samples with a deletion is still significantly higher in cases than WTCCC2 controls. Duplication burden remains relatively consistent across cohorts.

A CNV burden analysis aims to compare the burden of CNVs, defined as the total or average length of CNVs per sample, between a case and control sample. As with our binary analyses, we divided our analysis into cases vs. screened controls and cases vs. WTCCC2 controls. Commands used within PLINK can be found in the appendix. We calculated 1 and 2 sided p values using 10,000 null permutations of case-control status.

2.4.7.1 Cases Vs. Screened Controls

SUMMARY: Total CNV burden is significantly reduced in the screened control cohort, and in particular screened controls have a particularly low burden of deletion CNVs.

Here we present the results of PLINK burden analyses for our cases compared to screened controls, divided into analyses for all CNVs, deletion CNVs and duplication CNVs.

2.4.7.1.1 All CNVs

A significant difference is seen between cases and screened controls in the overall burden of CNVs with screened controls having a substantially lower overall burden of CNVs than cases.

CNV Type	**All CNVs**			
Analysis Cohorts	**Cases**	**Scr. Controls**	**1-sided p value**	**2-sided p value**
No. of CNVs	4298	321	N/A	N/A
No. of CNVs per sample	1.58	0.92	$<1.00x10^{-4}$	$<1.00x10^{-4}$
Proportion of samples having >0 CNVs	0.64	0.57	0.012	0.023
Total CNV burden per sample (kb)	563.5	400.4	0.00020	0.00070
Average CNV size per sample (kb)	243.7	250.5	0.71	0.66

Table 2.24. CNV burden figures for all CNVs with empirical 1 and 2-sided p values for 10,000 null permutations of case-control status.

2.4.7.1.2 Deletion CNVs

A particularly stark reduction in deletion CNV burden is seen when cases are compared to screened controls.

CNV Type	**Deletion CNVs**			
Analysis Cohorts	**Cases**	**Scr. Controls**	**1-sided p value**	**2-sided p value**
No. of CNVs	2702	144	N/A	N/A
No. of CNVs per sample	0.99	0.41	$<1.00x10^{-4}$	$<1.00x10^{-4}$
Proportion of samples having >0 CNVs	0.41	0.32	0.00020	0.00020
Total CNV burden per sample (kb)	475.7	275.7	$<1.00x10^{-4}$	$<1.00x10^{-4}$
Average CNV size per sample (kb)	213.8	213.1	0.58	0.58

Table 2.25. CNV burden figures for deletion CNVs with empirical 1 and 2-sided p values for 10,000 null permutations of case-control status.

2.4.7.1.2 Duplication CNVs

No significant differences in duplication CNV burden is observed between cases and screened controls.

CNV Type	Duplication CNVs			
Analysis Cohorts	**Cases**	**Scr. Controls**	**1-sided p value**	**2-sided p value**
No. of CNVs	1596	177	N/A	N/A
No. of CNVs per sample	0.59	0.51	0.10	0.19
Proportion of samples having >0 CNVs	0.39	0.38	0.45	0.86
Total CNV burden per sample (kb)	418.0	370.8	0.11	0.24
Average CNV size per sample (kb)	278.9	274.1	0.44	0.84

Table 2.26. CNV burden figures for duplication CNVs with empirical 1 and 2-sided p values for 10,000 null permutations of case-control status.

2.4.7.2 Cases Vs. WTCCC2 Controls

SUMMARY: Total CNV burden is notably similar between the two cohorts, although cases are again more likely to have deletion CNVs than controls via this analysis method.

Here we present the results of PLINK burden analyses for our cases compared to WTCCC2 controls, divided into analyses for all CNVs, deletion CNVs and duplication CNVs.

2.4.7.2.1 All CNVs

Within All CNVs a slight enrichment of CNVs is seen within the case cohort. Large sample numbers produce statistically significant results with relatively small case-control differences.

CNV Type	All CNVs			
Analysis Cohorts	**Cases**	**WTCCC2 Controls**	**1-sided p value**	**2-sided p value**
No. of CNVs	4298	6825	N/A	N/A
No. of CNVs per sample	1.58	1.41	0.0028	0.0040
Proportion of samples having >0 CNVs	0.64	0.61	0.015	0.034
Total CNV burden per sample (kb)	563.5	558.8	0.41	0.90
Average CNV size per sample (kb)	243.7	254.0	0.94	0.12

Table 2.27. CNV burden figures for all CNVs with empirical 1 and 2-sided p values for 10,000 null permutations of case-control status.

2.4.7.2.2 Deletion CNVs

A more pronounced enrichment of samples with deletion CNVs is seen in the case cohort when compared to the WTCCC2 cohort.

CNV Type	Deletion CNVs			
Analysis Cohorts	**Cases**	**WTCCC2 Controls**	**1-sided p value**	**2-sided p value**
No. of CNVs	2702	4063	N/A	N/A
No. of CNVs per sample	0.99	0.84	0.0033	0.0027
Proportion of samples having >0 CNVs	0.41	0.36	$<1.00\text{x}10^{-4}$	$<1.00\text{x}10^{-4}$
Total CNV burden per sample (kb)	475.7	464.8	0.33	0.66
Average CNV size per sample (kb)	213.8	223.2	0.87	0.28

Table 2.28. CNV burden figures for deletion CNVs with empirical 1 and 2-sided p values for 10,000 null permutations of case-control status.

2.4.7.2.3 Duplication CNVs

No significant differences are seen when duplication CNV burden is compared between cases and WTCCC2 controls.

CNV Type	Duplication CNVs			
Analysis Cohorts	**Cases**	**WTCCC2 Controls**	**1-sided p value**	**2-sided p value**
No. of CNVs	1596	2762	N/A	N/A
No. of CNVs per sample	0.59	0.57	0.31	0.60
Proportion of samples having >0 CNVs	0.39	0.38	0.25	0.48
Total CNV burden per sample (kb)	418.0	453.6	0.82	0.42
Average CNV size per sample (kb)	278.9	289.7	0.87	0.26

Table 2.29. CNV burden figures for duplication CNVs with empirical 1 and 2-sided p values for 10,000 null permutations of case-control status.

2.4.8 Singleton CNV Analysis

SUMMARY: We looked for significant association between singleton events, stratified by type, in our cases vs. the screened controls and the WTCCC2 controls. In both comparisons singleton deletion events are more common in cases than control cohorts, although some metrics are non-significant in the two comparisons.

Singleton CNVs are defined as occurring only once in a dataset, and are usually rare events. We compared the frequency of singleton events, stratified by type, in our case and control samples using PLINK (v1.07). PLINK commands for singleton analyses can be viewed in the appendix.

2.4.8.1 Cases Vs. Screened Controls

SUMMARY: The number of singleton CNVs per sample was significantly increased in cases when compared to screened controls, driven by singleton deletion CNVs. The total singleton deletion CNV burden was increased in cases when compared to screened controls. However the proportion of samples with a singleton CNV was not significantly different between cases and screened controls.

2.4.8.1.1 All Singleton CNVs

In the analysis for all singleton CNVs there was a significant difference in the number of CNVs per sample between cases and screened controls but no significant differences in other metrics.

CNV Type	**All Singleton CNVs**			
Analysis Cohorts	**Cases**	**Scr. Controls**	**1-sided p value**	**2-sided p value**
No. of CNVs	788	74	N/A	N/A
No. of CNVs per sample	0.29	0.21	0.017	0.047
Proportion of samples having >0 CNVs	0.21	0.19	0.15	0.29
Total CNV burden per sample (kb)	296	266.1	0.20	0.40
Average CNV size per sample (kb)	227.2	240.5	0.71	0.63

Table 2.30. CNV burden figures for all singleton CNVs with empirical 1 and 2-sided p values for 10,000 null permutations of case-control status.

2.4.8.1.2 Deletion CNVs

Within the analysis for singleton deletion CNVs there was a significant enrichment within cases, although again the average CNV size per sample was no

different between cohorts. In this analysis the total CNV burden per sample was significantly increased in cases compared to screened controls.

CNV Type	Singleton Deletion CNVs			
Analysis Cohorts	**Cases**	**Scr. Controls**	**1-sided p value**	**2-sided p value**
No. of CNVs	548	46	N/A	N/A
No. of CNVs per sample	0.20	0.13	0.017	0.044
Proportion of samples having >0 CNVs	0.14	0.13	0.29	0.52
Total CNV burden per sample (kb)	251.8	195.7	0.032	0.087
Average CNV size per sample (kb)	180.6	189.7	0.73	0.61

Table 2.31. CNV burden figures for singleton deletion CNVs with empirical 1 and 2-sided p values for 10,000 null permutations of case-control status.

2.4.8.1.2 Duplication CNVs

No significant differences in singleton duplication CNV burden was observed between cases and screened controls.

CNV Type	Singleton Duplication CNVs			
Analysis Cohorts	**Cases**	**Scr. Controls**	**1-sided p value**	**2-sided p value**
No. of CNVs	432	51	N/A	N/A
No. of CNVs per sample	0.16	0.15	0.37	0.68
Proportion of samples having >0 CNVs	0.13	0.14	0.65	0.81
Total CNV burden per sample (kb)	366.5	300.5	0.15	0.31
Average CNV size per sample (kb)	303.3	282.7	0.36	0.68

Table 2.32. CNV burden figures for singleton duplication CNVs with empirical 1 and 2-sided p values for 10,000 null permutations of case-control status.

2.4.8.2 Cases Vs. WTCCC2 Controls

SUMMARY: A significant enrichment for singleton deletion CNVs is seen within the case cohort when compared to the WTCCC2 control cohort. This is not observed for singleton duplication CNVs.

2.4.8.2.1 All Singleton CNVs

Within the analysis for all singleton CNVs a slight enrichment of singleton CNVs is seen within the case cohort, producing statistically significant results even with quite low absolute differences. Interestingly the total singleton CNV burden per sample and average singleton CNV burden per sample are broadly similar.

CNV Type	**All Singleton CNVs**			
Analysis Cohorts	**Cases**	**WTCCC2 Controls**	**1-sided p value**	**2-sided p value**
No. of CNVs	442	650	N/A	N/A
No. of CNVs per sample	0.16	0.13	0.0070	0.010
Proportion of samples having >0 CNVs	0.13	0.11	0.0054	0.0086
Total CNV burden per sample (kb)	269.8	257.7	0.25	0.52
Average CNV size per sample (kb)	221.8	212.1	0.22	0.45

Table 2.33. CNV burden figures for all singleton CNVs with empirical 1 and 2-sided p values for 10,000 null permutations of case-control status.

2.4.8.2.2 Singleton Deletion CNVs

When restricted to singleton deletion events a slightly more robust association is seen. Again, total and average singleton CNV burden per sample is not significantly different.

CNV Type	Singleton Deletion CNVs			
Analysis Cohorts	**Cases**	**WTCCC2 Controls**	**1-sided p value**	**2-sided p value**
No. of CNVs	341	465	N/A	N/A
No. of CNVs per sample	0.13	0.096	0.0020	0.0049
Proportion of samples having >0 CNVs	0.10	0.078	0.0010	0.0015
Total CNV burden per sample (kb)	220.9	226.4	0.63	0.72
Average CNV size per sample (kb)	176.2	189.3	0.87	0.28

Table 2.34. CNV burden figures for singleton deletion CNVs with empirical 1 and 2-sided p values for 10,000 null permutations of case-control status.

2.4.8.2.3 Singleton Duplication CNVs

No significant differences are seen across metrics when singleton duplication CNV burden is compared between cases and WTCCC2 controls.

CNV Type	Singleton Duplication CNVs			
Analysis Cohorts	**Cases**	**WTCCC2 Controls**	**1-sided p value**	**2-sided p value**
No. of CNVs	283	505	N/A	N/A
No. of CNVs per sample	0.10	0.10	0.54	0.96
Proportion of samples having >0 CNVs	0.086	0.086	0.53	1.00
Total CNV burden per sample (kb)	333.6	340.8	0.58	0.83
Average CNV size per sample (kb)	276.8	280.4	0.57	0.86

Table 2.35. CNV burden figures for singleton duplication CNVs with empirical 1 and 2-sided p values for 10,000 null permutations of case-control status.

2.4.9 Validation of CNVs Called by PennCNV

SUMMARY: We were able to follow up 40 CNVs, ranging in size from 106kb to 1.4MB, on an customised high density Agilent CGH array. All 40 CNVs validated, suggesting that we can have some confidence in the quality of our call set.

We identified 42 rare CNVs in 343 samples (291 cases and 52 screened controls) to follow up using a customised Agilent CGH array described in section 2.3.7. Of these 42 CNVs, 2 were present in samples where hybridisation to the array failed. Of the 40 CNVs left, all 40 were validated on the CGH array. The 40 CNVs were comprised of 28 deletions (length: mean=329,433bp, range=106,370-1,405,099bp and number of markers: mean=94, range 23-344) and 12 duplications (length: mean=481,894bp, range=126,731-1,394,315 and number of markers: mean=115, range=19-343). Figs. 2.26 and 2.27 illustrate the raw PennCNV calls, marker plots and CGH plots for the two CNV called with the least number of markers on the Illumina array (and therefore the least likely to replicate). Results are presented in Table 2.36.

Chr	CNV Start	CNV End	Length	CN	No. of Markers	Cohort	Valid-ated?
20	14,300,394	14,414,618	114,224	1	23	DeNT	Y
20	14,557,957	14,742,361	184,404	1	46	DeCC	Y
20	14,625,788	14,915,506	289,718	1	95	GENDEP	Y
20	14,679,595	14,785,965	106,370	1	48	GENDEP	Y
20	14,685,843	14,879,494	193,651	1	76	DeNT	Y
20	14,940,542	15,108,752	168,210	1	75	DeCC	Y
15	20,306,549	20,635,884	329,335	1	113	GENDEP	Y
15	20,306,549	20,687,235	380,686	1	115	DeCC	Y
15	20,321,135	20,629,449	308,314	1	108	GENDEP	Y
15	20,321,135	20,778,963	457,828	1	112	DeCC	Y
10	66,813,888	67,895,016	1,081,128	1	306	S'd Controls	Y
10	67,727,069	67,855,224	128,155	1	48	DeCC	Y
10	67,735,735	67,855,224	119,489	1	46	GENDEP	Y
10	67,885,161	68,114,740	229,579	1	74	DeCC	Y
10	67,927,636	68,139,483	211,847	1	67	GENDEP	Y
10	68,034,046	68,143,405	109,359	1	36	GENDEP	Y
10	68,064,862	68,194,973	130,111	1	47	DeNT	Y
10	68,426,851	68,579,571	152,720	3	46	DeCC	Y
10	68,998,527	69,125,258	126,731	3	19	DeNT	Y
3	7,566,405	7,755,116	188,711	1	63	DeNT	Y
3	7,999,019	8,253,215	254,196	3	60	GENDEP	Y
22	17,257,787	18,662,886	1,405,099	1	344	GENDEP	Y
22	17,292,678	18,686,993	1,394,315	3	343	DeCC	Y
22	17,426,677	17,578,226	151,549	3	40	DeNT	Y
22	19,066,315	19,792,353	726,038	1	174	GENDEP	Y
22	19,066,315	19,632,925	566,610	1	98	GENDEP	Y
22	21,328,337	21,979,242	650,905	1	121	GENDEP	Y
22	22,020,325	22,264,030	243,705	3	61	DeCC	Y
22	22,038,833	23,326,630	1,287,797	3	241	DeCC	Y
22	22,417,319	23,326,630	909,311	3	149	DeCC	Y
6	161,239,662	161,706,684	467,022	3	135	GENDEP	Y
6	162,264,702	162,677,104	412,402	3	136	DeCC	Y
6	162,269,448	162,385,193	115,745	1	28	DeNT	Y
6	162,447,621	162,649,371	201,750	1	70	DeNT	Y
6	162,562,714	162,921,073	358,359	1	128	DeCC	Y
6	162,610,624	162,809,965	199,341	1	81	DeNT	Y
6	162,637,688	162,834,976	197,288	3	77	DeNT	Y
6	162,644,237	162,829,925	185,688	3	75	DeNT	Y
6	162,767,020	162,903,833	136,813	1	59	DeCC	Y
6	162,824,155	162,956,501	132,346	1	40	DeNT	Y

Table 2.36. 40 CNVs out of 40 called by PennCNV, of a variety of sizes and copy number states, validate on a customised CGH array.

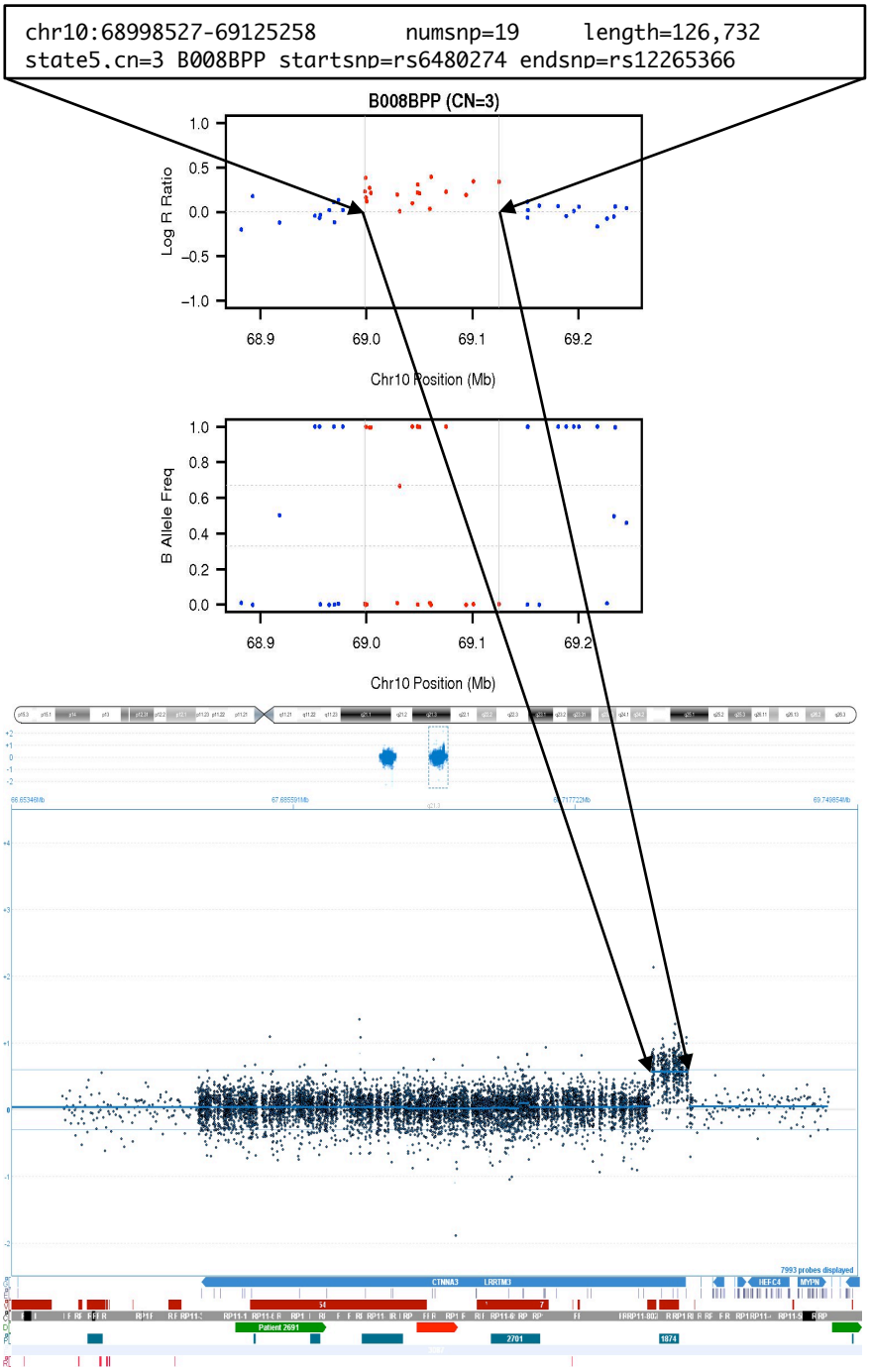


Fig 2.26. A small duplication CNV on chromosome 10 validates on a high density CGH array.

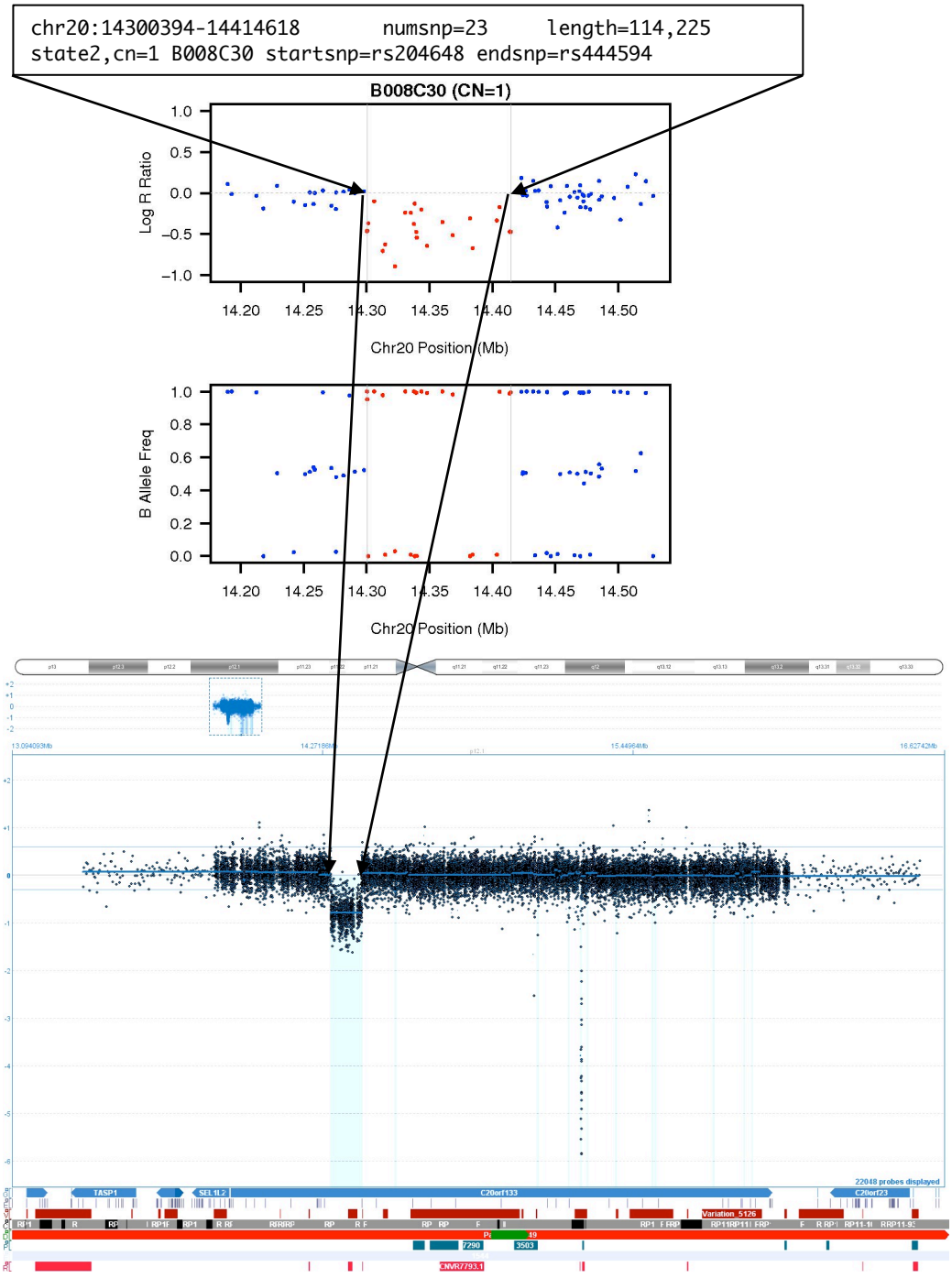


Fig. 2.27. A small deletion CNV on chromosome 20 validates on a high density CGH array.

2.5 Conclusion

We called CNVs in over 10,000 samples of DNA derived from a mixture of venous blood (cases, screened controls, WT2 controls), cell lines (WT2 controls) and cheek swabs (screened controls) using a consensus marker set derived from the Illumina 610 Quad and 1M arrays using PennCNV. We found that samples derived from cheek swabs were systematically different from samples derived from venous blood and cell lines, and excluded them from our analysis. We removed poorly performing samples and calls more likely to be false positive and then compared the frequency of samples with different types of CNV in our case and control cohorts. We observed a statistically significant increase in frequency of samples harbouring a rare deletion CNV from the screened controls, through our WTCCC2 controls with the highest frequency in the case cohort. Singleton and burden analyses also reinforced this finding. To account for poorly performing samples and population stratification we performed two post hoc analyses removing the worst performing 10% of samples across cohorts and restricting our cohort to a UK-only sample set. We found that our initial association remained robust although some poor quality samples in our original analysis were likely to be contributing false positive calls which may have inflated our statistical association.

Chapter 3. Common Copy Number Variants

3.1 Introduction

Common copy number variants (CNVs) are often called copy number polymorphisms (CNPs). They have been arbitrarily defined by different groups as occurring in more than 10% of a given population(Conrad et al., 2010) and more than 5% of a given population (Craddock et al., 2010), with figures of more than 1% of a given population often being used in studies of rare CNVs(Grozeva et al., 2010; McQuillin et al., 2011; Rucker et al., 2011). The frequency of copy number polymorphisms varies between populations, and therefore what is and is not considered to be a polymorphism will vary dependent on the sample being studied and the threshold that is chosen§. Some have argued that common variants account for most of disease susceptibility(Risch & Merikangas, 1996), others that they account for little(Conrad et al., 2010) and others that a mixture of liabilities probably exist(Uher, 2009). Craddock et al. noted that some single nucleotide polymorphisms (SNPs) reliably tag CNPs. That is, a given allele in a given SNP segregates with the presence or absence of a CNP. They provided a reference dataset within this work describing 3,432 CNPs tagged by SNP.

3.2 Hypotheses

We aimed in this experiment to determine whether

A) CNPs tagged by SNPs within our consensus marker set described in chapter 2 were associated with caseness in our samples.

3.3 Methods

SUMMARY: We used a UK-only sample set in this analysis to avoid errors due to population stratification, including control samples derived from cheek swabs that we had excluded in our rare CNV analysis. Individual samples and SNP genotypes were cleaned according to well-established principles in SNP genome wide association study (GWAS) analysis. From a cleaned SNP marker set we identified 516 SNPs tagging CNPs that had previously been identified by Craddock et al(Craddock et al., 2010), which we tested for association with caseness, correcting for two principle components derived from the program Eigenstrat, using PLINK v1.07(Purcell et al., 2007).

3.3.1 Samples

3.3.1.2 Cases

We drew our samples from the DeCC, DeNT and GENDEP samples described in chapter 2, restricting our sample set to those with UK origin. 1,346 cases (69.3% female) were used from the DeCC sample, 332 cases (75.3% female) from the DeNT sample and 88 (63.6% female) from the GENDEP sample for a total of 1,766 cases.

3.3.1.3 Controls

We compared our cases to 1,989 screened control samples taken from the DeCC and BaCC samples, restricted to those of UK origin, and 5,069 population control samples from phase 2 of the Wellcome Trust Case Control Consortium, collected as described in chapter 2. For this analysis we used control samples derived

from cheek swab DNA, which we had excluded in our rare CNV analysis. The rationale for this is that within a tagSNP analysis, which relies purely on biallelic SNPs, as opposed to SNPs which may exist in regions of CNV and therefore be multiallelic, it is possible to exclude poorly performing markers on the basis of various metrics assuming a biallelic genotype (e.g. Hardy-Weinberg equilibrium). Thus SNPs in regions affected by copy number will be excluded, and given that aberrant copy number variation in specific regions of the genome within these samples was the basis for the concern with our cheek swab samples, they can now be included.

3.3.2 Genotyping and Quality Control

Samples were collected and DNA extracted and genotyped as described in chapter 2. In this instance, the CNG laboratory supplied us with genotype files with SNP allele calls. SNP genotype data from the WTCCC2 controls was downloaded from the European Bioinformatics Institute (http://www.ebi.ac.uk/).

We used PLINK(Purcell et al., 2007) for initial quality control of the entire dataset, before restricting our marker set to those tagging CNPs. Quality control was based on the paper by Lewis et al.(Lewis et al., 2010). We initially excluded samples with a genotype call rate of less than 95% or with outlier values for the proportion of heterozygous allele calls (<0.29 or >0.36). Individuals were also excluded if the phenotypic gender did not match the sex assigned by the SNP analysis, or was inconclusive. Related and duplicate individuals within and across case and control samples were identified through identity-by-state

sharing analysis on a linkage disequilibrium-pruned set of SNPs (N=~18,000). If individuals were found to be related to any other study member (up to second degree), or duplicates, then the sample with the lowest genotype call rate was excluded. Non-European ancestry was determined using Eigenstrat (see below).

SNP quality control methods were applied separately to each cohort before merged datasets were created. As discussed in chapter 2, a large proportion of our screened control samples were derived from cheek swabs, and thus the genotype call rate was lower in these samples than in samples derived from blood. Thus genotype call rate thresholds were applied separately in each group. SNPs with a call rate of less that 99%, that diverged from Hardy-Weinburg equilibrium at a p value of less than $1x10^{-5}$, or with a minor allele frequency of less than 1%, were removed. Based on this high quality SNP set, samples with a genotype call rate of less than 99% were then removed. After quality control 1,628 cases, 1,588 screened controls, 4,566 WTCCC2 controls and 541,628 SNPs were included for onward analysis.

EIGENSTRAT software was used to analyse all samples for population stratification using principal components(Price et al., 2006). We used 80,304 SNPs, omitting regions with high levels of linkage disequilibrium. We then identified individuals with non European ancestry by combining our genotypes with those derived from HapMap data from Utah residents with ancestry from northern and western Europe (CEU), Yoruba in Ibadan, Nigeria (YRI), Han Chinese in Beijing, China (CHB), Japanese in Tokyo, Japan (JPT), and Gujurati Indians in Houston, Texas (GIH) populations. GIH were included since much non-

European ancestry in the UK is from the Indian sub-continent. Samples with non-European ancestry were omitted.

After removal of non-European Caucasian outliers, Eigenstrat was re-run without the HapMap populations to derive principal components for inclusion in the analysis. The UK-ascertained cases provided a good match to control samples, and two principal components were necessary to correct for differences between cases and controls.

We then took the table of 3,432 CNPs tagged by SNPs (tagSNPs) from the paper by Craddock et al(Craddock et al., 2010), and restricted our choice of possible tagSNPs to those contained within our QC'd marker set and those with a square of the correlation coefficient between SNP and CNP (r^2) value of greater than 0.8. r^2 is a metric calculated to represent the degree of correlation between two alleles, in this case a SNP and a CNP. 516 SNPs (mean r^2 = 0.970, SD 0.0451, Fig. 3.1) were used in our analysis. Most SNPs tag CNPs almost perfectly (R^2>0.99). A full list of SNPs is shown in the appendix.

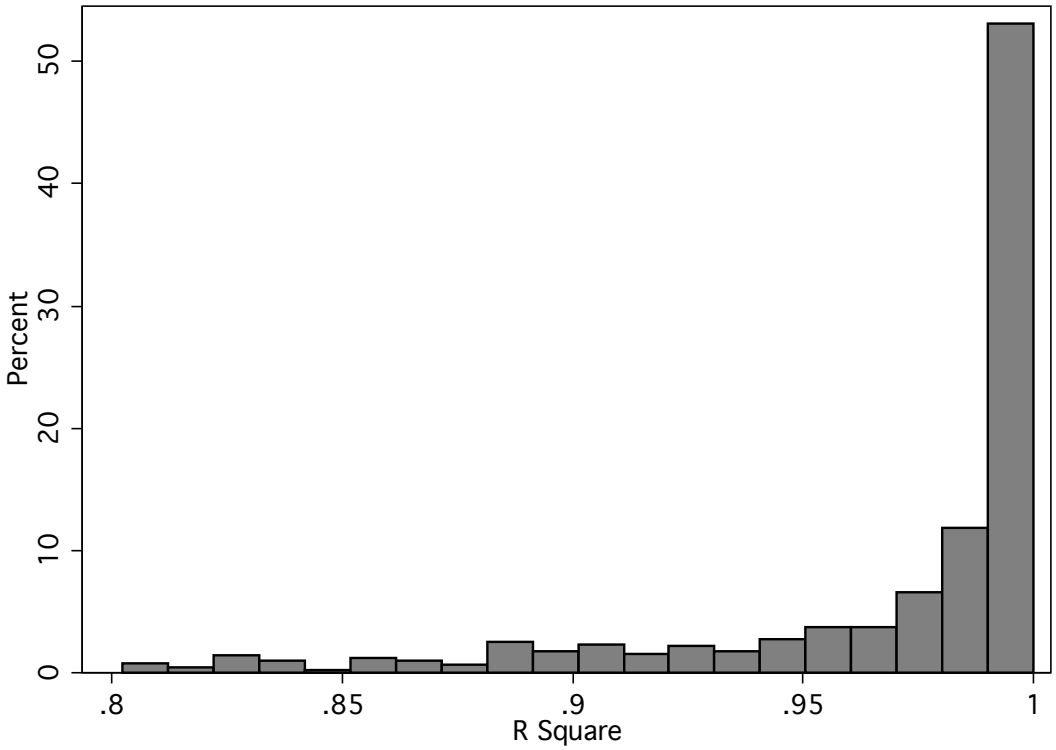


Fig. 3.1. Histogram of 516 SNP r^2 values used in our CNP analysis.

3.3.3 Association Testing

Principal component loadings differed between cases and controls on the first two principal components, and these were therefore included as covariates in testing for association. We used PLINK v1.07 to test all 516 SNPs for association with caseness using a logistic regression model and adaptive permutation to calculate p values in our cases vs. our screened controls and our cases vs. the WTCCC2 controls. PLINK commands can be found in the appendix. Permuted p values for significance are reported, although we found that the difference between the two methods was minimal. We set a Bonferroni-corrected p value of $0.05/516 = 9.69 \times 10^{-5}$ corresponding to a p-value for significance of 1 test of 0.05.

3.4 Results

SUMMARY: No tagSNP achieved genome-wide significance after correction for multiple testing. One tagSNP (rs12035407) in our cases vs screened control sample analysis replicated in our cases vs WTCCC2 analysis with a $p < 0.05$. This SNP tags a deletion CNV over a gene that has not been implicated in psychiatric disorders.

3.4.1 Cases vs. Screened Controls

After QC, 1,628 cases and 1,588 screened controls were used in our analysis. A QQ plot (Fig. 3.2) of observed p values versus expected p values suggests no systematic bias in our analysis after correction for two principal components. Genomic lambda, mean = 1.05, median = 1.01

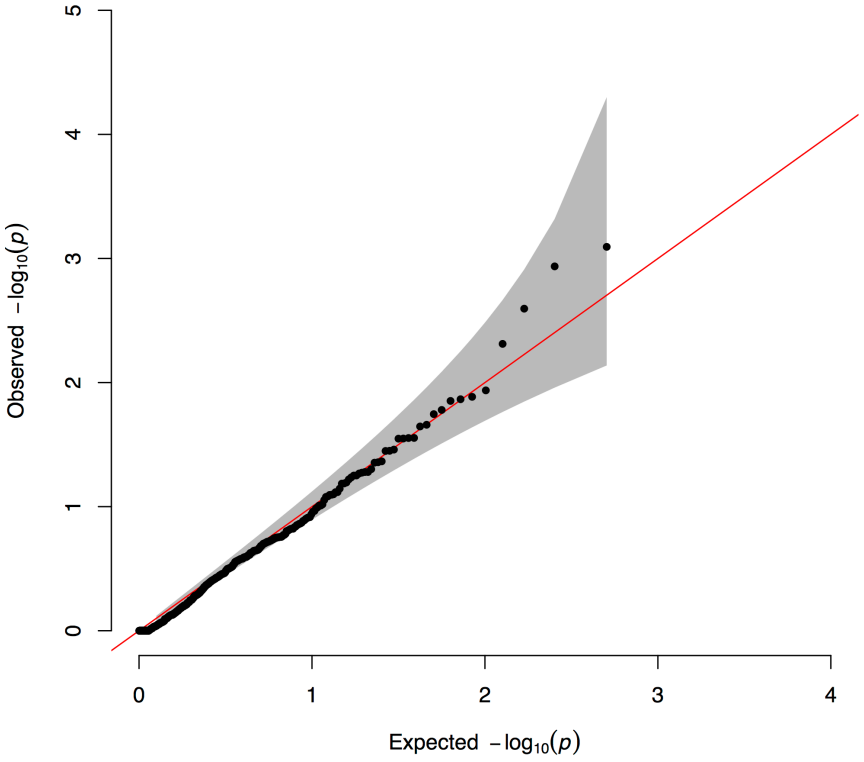


Fig. 3.2. QQ plot of observed vs. expected p values in our association study. Shaded area indicates 95% confidence interval.

A manhattan plot (Fig. 3.3) illustrates all tagSNPs by chromosome and position along with the inverse log of the p value for association. No tagSNP p value survives Bonferroni correction for multiple testing (9.7×10^{-5}).

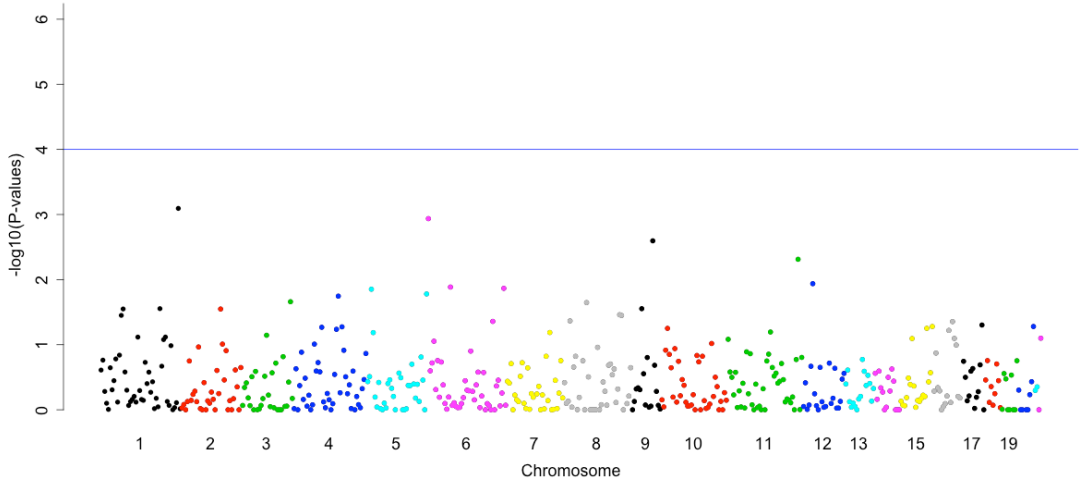


Fig. 3.3. Manhattan plot of -$\log_{10}$ p-values for association assuming a corrected p level of significance of 9.7×10^{-5} (blue line).

Table 3.1 denotes the top five tagSNPs along with the CNP they tag.

SNP	Location	P value	OR (95% CI)	N Cases with CNP (%) | N Controls with CNP (%)	Gene
rs12035407	chr1: 238,445,281	7.5×10^{-4}	1.22 (1.09-1.37)	422 (25.8%) | 377 (23.8%)	FMN2
rs9405611	chr6: 3,089,027	0.0013	0.84 (0.75-0.93)	459 (28.2%) | 484 (30.5%)	BPHL
rs2174926	chr9: 120,565,371	0.0034	1.17 (1.06-1.29)	785 (48.2%) | 730 (46.0%)	None
rs540029	chr11: 121,089,627	0.0043	1.19 (1.01-1.35)	378 (23.2%) | 349 (22.0%)	None
rs2564577	chr12: 30,290,692	0.012	1.14 (1.03-1.23)	664 (40.8%) | 602 (37.9%)	None

Table 3.1. Top 5 tagSNPs in our CNP association study comparing cases to screened controls.

Our top associated tagSNP was rs12035407 ($p=7.5 \times 10^{-4}$, OR 1.22 (95% CI 1.09-1.37)). This SNP tags ($r_2=1.00$) a 1,596 bp deletion CNP (chr1:238,459,885-238,461,481) over the gene FMN2 with a minor allele frequency (MAF) of 0.243(Craddock et al., 2010). FMN2 (formin 2) is one of a family of formin homology domain proteins which play a role in cytoskeletal organization and the establishment of cell polarity. This gene has not been associated with

neuropsychiatric disorders. The SNP is 14.6kb from the start of the CNP. Insufficient markers from our consensus marker set or the full 610 Quad marker set cover the CNP region in question so we could not follow up the frequency of the deletion CNP in our PennCNV calls. Indeed this was the case for all the top five CNPs in our analysis. In our subsequent comparison of our cases to the WTCCC2 controls this tagSNP demonstrated a p value for association of 0.037 (OR 1.10 (95%CI 1.01 - 1.21)).

3.4.2 Cases vs. WTCCC2 Controls

After QC, 1,628 cases and 4,566 WTCCC2 controls were used in our analysis. A QQ plot (Fig. 3.4) of observed p values versus expected p values suggests no systematic bias in our analysis. Genomic lambda, mean = 1.10, median = 1.05.

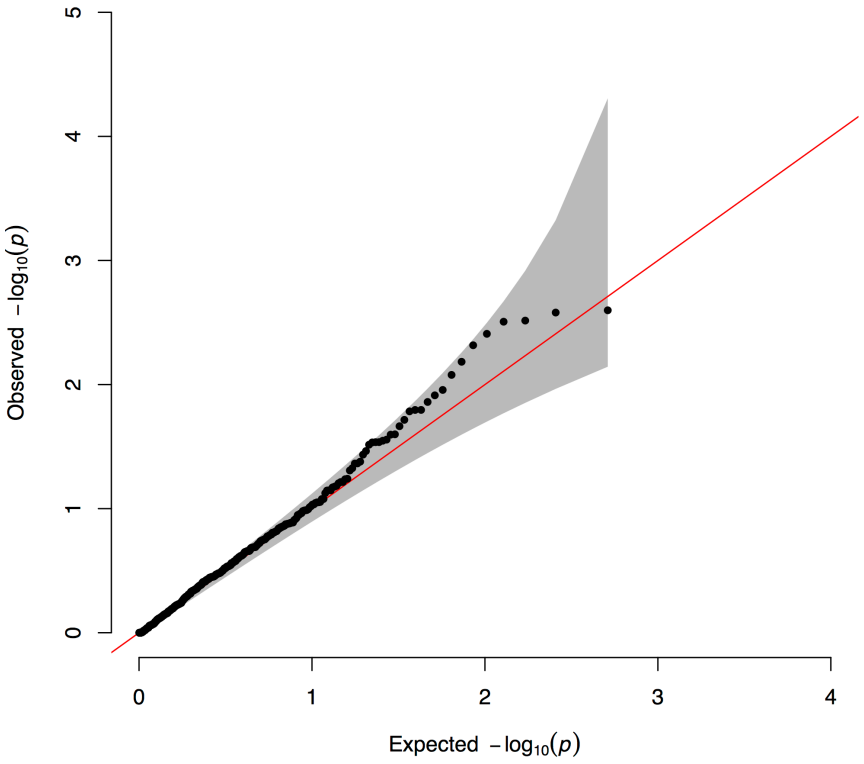
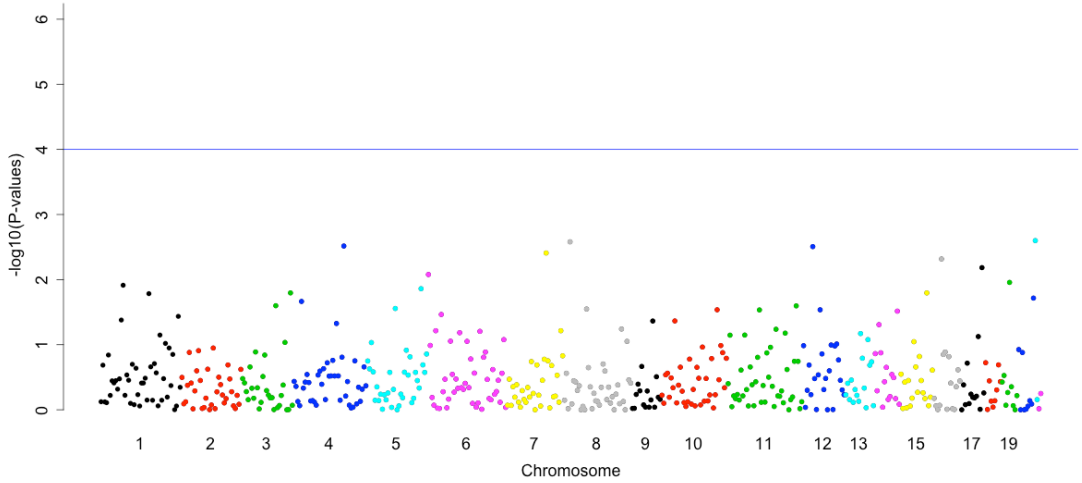


Fig. 3.4. QQ plot of observed vs. expected p values in our association study. Shaded area represents 95% confidence interval.

A manhattan plot (Fig. 3.5) illustrates all tagSNPs by chromosome and position along with the inverse log of the p value for association. No tagSNP p value survives Bonferroni correction for multiple testing ($9.7x10^{-5}$).



(Parker et al., 2000)(Parker et al., 2000)Fig. 3.5. Manhattan plot of -$\log_{10}$ p-values for association assuming a corrected p level of significance of 9.7×10^{-5} (blue line).

Table 3.2 denotes the top five tagSNPs along with the CNP they tag.

SNP	Location	P value	OR (95% CI)	N Cases with CNP (%) | N Controls with CNP (%)	Gene
rs13046557	chr21: 15,510,230	0.0024	0.88 (0.81-0.96)	716 (44.0%) | 2155 (47.2%)	None
rs17326768	chr8: 3,567,534	0.0026	0.88 (0.81-0.96)	597 (36.7%) | 1804 (39.5%)	CSMD1
rs11100904	chr4: 147,123,754	0.0037	0.88 (0.81-0.96)	638 (39.2%) | 1941 (42.5%)	None
rs2564577	chr12: 30,290,692	0.0040	1.13 (1.04-1.23)	664 (40.8%) | 1731 (37.9%)	None
rs370760	chr7: 101,501,942	0.0040	0.85 (0.77-0.95)	267 (16.4%) | 854 (18.7%)	CUX1

Table 3.2. Top 5 tagSNPs in our association study comparing cases to WTCCC2 controls.

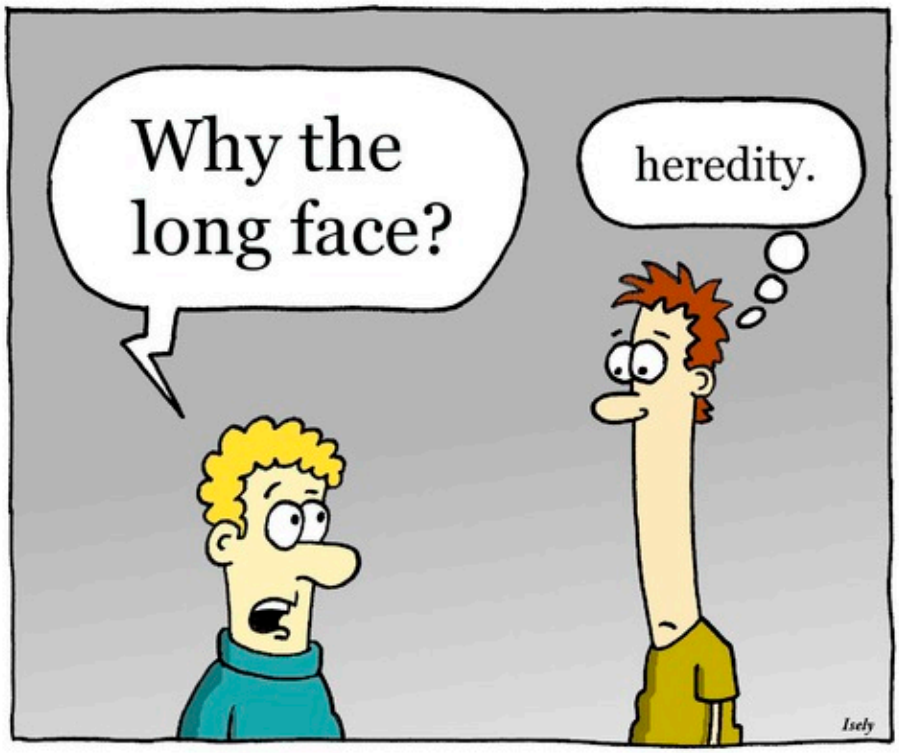
The SNPs with the lowest p values (0.0024 & 0.0026 respectively) were rs13046557, located on chromosome 21 (chr21:15,510,230), and rs17326768, located on chromosome 8 (chr8:3,567,534). Rs13046557 tags a 1.5kb duplication CNP with a minor allele frequency of 0.296, and is located approximately 200kb downstream of the gene NRIP1. Our analysis suggests that the duplication CNP confers a protective effect. NRIP1, or nuclear receptor

interacting protein 1 is a nuclear protein that specifically interacts with the hormone-dependent activation domain AF2 of nuclear receptors and is highly expressed in the pre-frontal cortex. Also known as RIP140, this protein modulates transcriptional activity of the estrogen receptor. Female mice null for this gene are infertile (Parker et al., 2000). The CNP is approximately 600kb upstream of the gene USP25. USP25 or ubiquitin specific protease 25, is part of a family of proteins which mediate release of ubiquitin release from degraded proteins by disassembly of polyubiquitin chains. Ubiquitin is a protein involved in regulation of intracellular protein breakdown, cell cycle regulation, and the stress response (Valero et al., 1999). Whilst both these genes are relatively interesting, in the previous analysis comparing our cases to screened controls this tagSNP does not replicate the degree of association (p=0.50 (OR 0.97, 95% CI 0.88-1.07)). Insufficient markers were available on all of the top five CNPs to follow up our tagSNP CNP calls with our PennCNV calls.

3.5 Conclusion

In an analysis of common copy number variants, often called copy number polymorphisms, we used information from previous research providing a reference of SNPs that tag CNPs. Accounting for population stratification, both by restricting our analysis to a UK-only sample set, and also by correcting in our analysis for principal components derived using the program Eigenstrat(Price et al., 2006), we were able to show that none of the CNPs tagged by well-performing SNPs in our consensus marker set were associated when our cases were compared to screened controls or WTCCC2 controls.

Chapter 4. Chromosome 22q11.2



4.1 Introduction

In this fourth chapter we will describe a more in-depth analysis of copy number variation in chromosome 22q11.2 and follow up with a high density oligonucleotide comparative genomic hybridisation (CGH) array in a selected cohort of samples. This region deserves special attention, as it has been robustly associated with neuropsychiatric disorders, harbours numerous genes known to play a role in neuronal development and functioning and, of all the regions of the human genome, is the most prone to rearrangement(Shaikh et al., 2000).

4.2 The Evolution and Function of 22q11.2

Several chromosomal regions in the human genome, including 22q11, are notable for the presence of low-copy repeats (LCRs), also known as segmental duplications(Shaikh et al., 2000). Segmental duplications are regions of the genome from a few to several hundred kilobases in length that share high (usually >95%) sequence similarity (homology) and can include genomic sequences from the introns and exons of existing genes(Eichler, 2001). These regions occur because of evolutionary mechanisms that predispose regions of the genome, including genes, to duplicate. On the one hand this allows genes to diverge and evolve different functions without the need for the gene to be re-evolved from scratch (in this sense, it is Nature's way of not needing to 're-invent the wheel'). On the other hand it provides a mechanism to increase the chances of certain regions of the genome (presumably those under active selection by evolution) rearranging. Rearrangement tends to occur when segmental duplications flank certain regions of the genome(J. A. Bailey, Yavor, Massa, Trask,

& Eichler, 2001). The sequence similarity between the flanking segmental duplications means that they are more likely than chance to be mismatched during mitosis or meiosis, which then results in inter or intra chromosomal rearrangements, deletions or duplications of the region flanked. Chromosome 22q11.2 is unique within the genome, because the segmental duplications contained within it contain unusual palindromic AT-rich repeats and duplications of the breakpoint cluster region (BCR) gene, that appear to further increase the rate of both random and recurrent variation(Emanuel, 2008). In essence, 22q11.2 is a genetic melting pot in the human genome, and, coupled with the fact that many of the genes within this region are highly expressed in the developing and mature brain, this makes this region a highly pertinent area to study in psychiatric disorders.

4.3 22q11.2 Deletion Syndrome

Chromosome 22q11.2 is home to the most frequently occurring deletion syndrome in the human genome, a 3MB deletion that is associated with a variety of congenital and neurodevelopmental abnormalities, resulting in eponymous syndromes such as DiGeorge syndrome (DGS), Velocardiofacial syndrome (VCFS) and conotruncal anomaly face syndromes (CAFS)(Robin & Shprintzen, 2005). The 22q11.2 duplication syndrome is also described, and in keeping with many reciprocal syndromes of deletion syndromes, it has a much milder phenotype(Yobb et al., 2005). Depending on how one defines the region and the genes present, there are about 60 genes within the 22q11.2 region, and many are associated with neuronal development and function.

Those with the 22q11.2 deletion have a variety of different types and severity of physical and cognitive impairments and a striking minority (far more than would be expected by chance) go on to develop neuropsychiatric disorders, especially schizophrenia(Karayiorgou et al., 2010), ADHD(Niklasson & Gillberg, 2010) and mood disorders(Jolin et al., 2009; Papolos et al., 1996). A study of 22q11.2 mouse orthologues found that many genes within the 22q11.2 region are expressed in the developing mouse forebrain, and that some are expressed in adult brain in regions implicated in schizophrenia in humans(Maynard et al., 2003).

Given the unique and on-going evolutionary profile of 22q11.2, its high concentration of genes involved with neurodevelopment and function and its robust association with neuropsychiatric disorders, we decided to more intensively study our PennCNV calls in this area and follow up our calls by designing a high density customised oligonucleotide comparative genomic hybridisation microarray to characterise the region at a higher resolution in a subset of our samples. Comparative genomic hybridisation (CGH) involves labelling a test DNA sample with a fluorescent dye and a reference DNA sample of known copy number with a different fluorescent dye and hybridising both samples together to a genomic probe of known sequence (Kallioniemi et al., 1992). The ratio of fluorescence between the test DNA sample and the reference DNA sample can then be used to infer the relative copy number in the test sample. This technique has been miniaturised to allow thousands or millions of CGH experiments to take place on a single glass slide, using technologies not dissimilar to those used to produce GWAS microarray chips(Pinkel et al., 1998). The resulting technology is termed ‘array CGH’ (aCGH).

4.2 Hypotheses

A) our cases will have an increased frequency of CNVs, particularly deletion CNVs, in the 22q11.2 region, given its prior association with psychiatric disorders.

B) this increased frequency will be reflected in a subset of cases and controls followed up with a high density oligonucleotide CGH array

4.3 Methods

4.3.1 Selection of Samples for Follow Up

Sample collection and genotyping methods have already been described in chapter 2. We only followed up cases and screened controls because we had only been granted access to the WTCCC2 GWAS chip data, not DNA samples.

Since we were using only samples genotyped on the 610 Quad array (having excluded the WTCCC2 controls), we re-analysed our raw data with PennCNV, but this time using all markers (n=620,901) rather than the consensus marker set described in chapter 2. In processing our call set we used identical methods described in 2.3.5.4 except that we restricted our calls to those over the 22q11.2 region, we did not exclude CNVs by size, we lowered our threshold for CNVs made with consecutive SNPs from 10 to 5 and did not exclude common CNVs.

From our PennCNV calls we identified 816 samples (82 screened controls and 734 cases) with calls over the 22q11.2 region that could theoretically be followed up. Within these 816 samples were a total of 1080 CNV calls (367

deletions and 713 duplications). 325 samples (274 cases and 51 screened controls) with CNVs called by PennCNV were randomly selected and followed up with the array. A subset of 207 randomly selected PennCNV calls from these 325 samples were then validated with the array CGH data.

Array CGH genotyping methodology can be found in chapter 2.3.7. The sample numbers quoted here differ from those quoted in 2.4.9 because in this section we are only describing samples selected on the basis of the presence of 22q11.2 calls, rather than other regions we followed up with the array.

52,476 probes covered the 22q11.2 region (chr22:16,300,001-24,300,000), which itself is approximately 8MB in size, equating to an average coverage density of 1 probe every 152 base pairs. Regions of segmental duplication, by definition, are not covered routinely by oligonucleotide arrays as probes do not reliably map to unique genomic segments in these areas.

4.3.4 Data Analysis

4.3.4.1 PennCNV Data

To assess our PennCNV calls across the 22q11.2 region we used our calls described in chapter 2, however modified our call set as described above and restricted our calls to any CNV falling over any part of the 22q11.2 region (chr22:16,300,001-24,300,000). We restricted our sample set as described above and re-analysed our data using PLINK v1.07, stratifying our data into deletions, duplications, common CNVs (>1% of total sample frequency) and rare CNVs (<1% of total sample frequency). We used PLINK's --mperm and --cnv-test-

2sided function to perform 10,000 null permutations of case control status to calculate empirical 1-sided and 2-sided p values (as it may be reasonable to expect that controls have more of some events than cases). We calculated statistical significance of the difference between the rate of CNVs per subject, the proportion of subjects to have at least one CNV, the total CNV distance spanned per subject and the average CNV size per subject. Scripts can be found in the appendix.

4.3.4.2 aCGH Data

To process CGH data we used DNAcopy(Olshen, Venkatraman, Lucito, & Wigler, 2004), a well-established package for the analysis of array CGH data that uses circular binary segmentation. Circular binary segmentation is a modification of binary segmentation(Sen & Srivastava, 1975) that takes better account of the fact that true deviations in copy number may not accurately be reflected by all the contiguous probes involved in detecting that change, usually due to random noise. DNAcopy is implemented within the R programming language(R Development Core Team, 2005). We processed raw Log_2 ratio data (sample dye fluorescence intensity as a ratio of reference dye fluorescence intensity) for each probe in each sample from the data provided by OGT. We used the standard protocol provided by the authors of DNAcopy to normalise, smooth, segment and plot our data (see appendix for scripts).

To determine thresholds for calling a deletion or duplication we initially plotted a histogram of segment means taken from segments called over autosomal regions by the algorithm with more than 100 consecutive probes (and thus likely

to represent robustly segmented regions of similar copy number) and then visually looked at a variety of plots of $\log_2$ ratios to determine a threshold for further analysis of an appropriate deletion and duplication threshold (Fig. 4.1, upper panel). As can be seen in the upper panel of Fig. 4.1, the vast majority of segment means fall around 0, representing regions of normal copy number. We then experimented with excluding segments with different thresholds and re-plotting the histogram to achieve a bimodal distribution which most likely represented the two clusters of segment means representing deletion and duplication events. The distribution appeared most convincing when excluding segment means with a value of more than -0.1 and less than 0.1 (Fig. 4.1, lower panel).

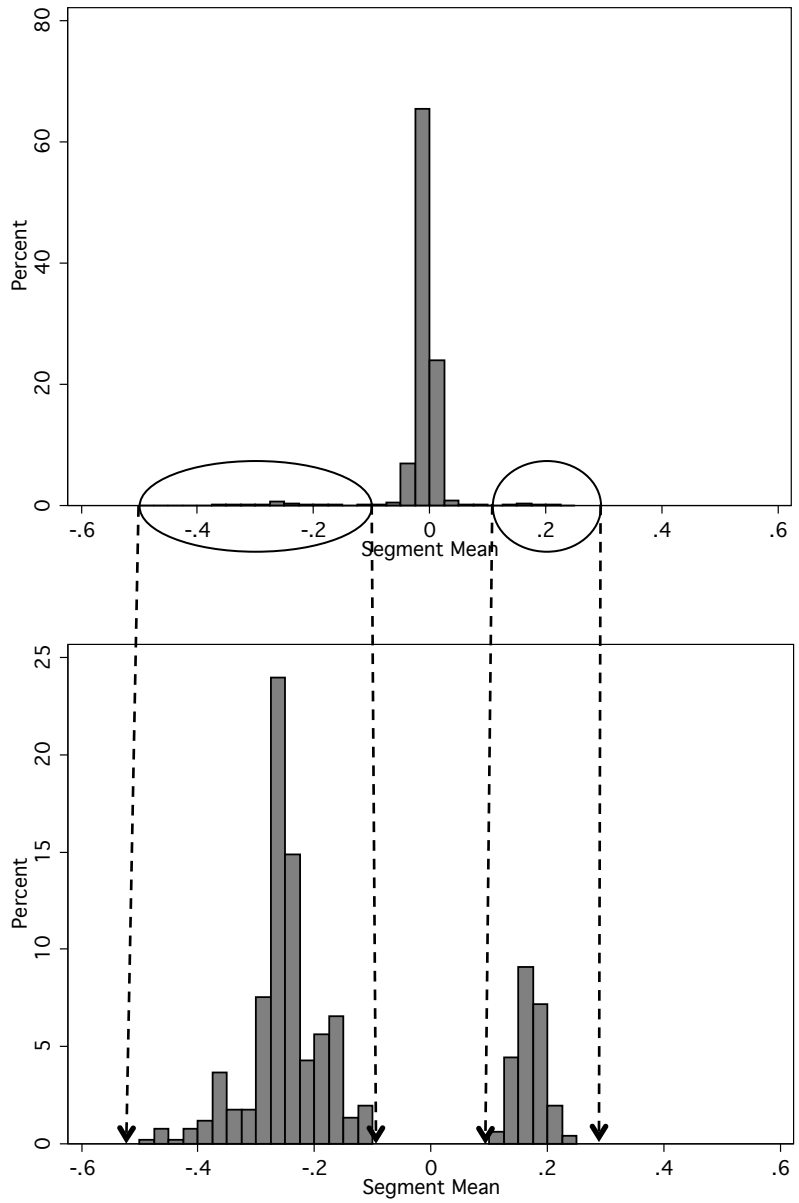


Fig. 4.1 Histogram of segment means for aCGH data. All data (top graph) and with values falling between -0.1 and 0.1 removed (lower graph) to show distributions for segment means of deletions (left) and duplications (right).

We recalculated the mean and standard deviations for the remaining segment means, stratified into groups representing putative deletions (n=395) and duplications (n=134). This resulted in 395 deletion segments and 134 duplication segments with a mean and standard deviation (SD) of -0.25 & 0.064

and 0.16 & 0.033 respectively. We set our thresholds based on 2xSD subtracted from the mean, resulting in a threshold segment mean of -0.124 for deletions and 0.0957 for duplications. We then only considered segments within the 22q11.2 region called with 10 or more contiguous markers in our analyses.

To visualise our array CGH data we used OGT's supplied Cytosure Interpret software, which plots normalised Log_2 ratios along an X axis of chromosomal position and allows the user to zoom in and out of the data and save representative images of calls easily. We used this feature to manually follow up the calls made by PennCNV on our Illumina 610 Quad data in cases and screened controls with which we had array CGH data.

4.4 Results

4.4.1 Sub-analysis of 22q11.2 Using PennCNV Calls from Illumina 610 Quad Data

SUMMARY: The proportion of samples with rare deletions is significantly increased in cases compared to controls. There is also evidence that the average total distance of rare duplications are significantly increased in cases compared to controls. Results for common variants have not been controlled for effects due to population stratification and should be viewed with caution, however there are no significant differences across our analyses.

Table 4.1 illustrates the CNV event rate per person. The rate of CNV deletion segments is significantly increased in cases when compared to controls, and this is principally driven by rare deletions as opposed to common deletions. This is similar to the results of our analysis across the genome shown in chapters 2 and 3. Interestingly there is also a driving effect from rare duplications which fails to reach significance.

Event Frequency	Event Type	Rate (cases | controls)	1 sided p value	2 sided p value
All	Dels/Dups	0.40 | 0.35	0.13	0.25
All	Dels	0.14 | 0.078	**0.018**	0.055
All	Dups	0.26 | 0.27	0.64	0.82
Common	Dels/Dups	0.31 | 0.33	0.73	0.62
Common	Dels	0.073 | 0.071	0.53	1.00
Common	Dups	0.22 | 0.26	0.90	0.22
Rare	Dels/Dups	0.096 | 0.021	**0.0005**	**0.0043**
Rare	Dels	0.084 | 0.014	**0.0004**	**0.0056**
Rare	Dups	0.039 | 0.014	0.055	0.13

Table 4.1. 22q11.2 analysis of PennCNV GWAS calls. CNV event rate (Rate) per person. Significance values of less than 0.05 are highlighted in bold.

Table 4.2 illustrates the proportion of cases and controls to have at least one event (this is similar in nature to our analysis in chapter 2). Again, rare deletion events are significantly more common in cases than screened controls and there appears to be a non-significant driving effect from duplications.

Event Frequency	**Event Type**	**Prop (cases | controls)**	**1 sided p value**	**2 sided p value**
All	Dels/Dups	0.30 | 0.29	0.41	0.79
All	Dels	0.096 | 0.075	0.15	0.28
All	Dups	0.22 | 0.25	0.82	0.41
Common	Dels/Dups	0.26 | 0.28	0.80	0.48
Common	Dels	0.064 | 0.071	0.72	0.70
Common	Dups	0.20 | 0.24	0.93	0.19
Rare	Dels/Dups	0.065 | 0.021	**0.001**	**0.004**
Rare	Dels	0.055 | 0.014	**0.0014**	**0.0035**
Rare	Dups	0.029 | 0.014	0.096	0.18

Table 4.2. 22q11.2 analysis of PennCNV GWAS calls. Proportion of cases/controls to have at least one event (Prop). Significance values of less than 0.05 are highlighted in bold.

Table 4.3 illustrates figures for the total event size per subject. Here we see that the total event size per subject for duplications overall, and principally driven by rare duplications, is significantly greater in cases than controls. This effect is driven by a number of large, rare duplications in our cases and a relative absence of such events in our screened controls. Interestingly there is no significant association with deletions, suggesting that if a subject has a deletion event, then it is no more likely to be larger in a case than a control.

Event Frequency	Event Type	KbTot (cases | controls)	1 sided p value	2 sided p value
All	Dels/Dups	115 | 90.74	0.077	0.20
All	Dels	109.2 | 118.6	0.73	0.74
All	Dups	108 | 71.72	**0.011**	0.073
Common	Dels/Dups	92.36 | 87.68	0.38	0.71
Common	Dels	87.47 | 110.9	0.85	0.30
Common	Dups	89.25 | 71.6	0.092	0.21
Rare	Dels/Dups	176.3 | 84.08	0.092	0.23
Rare	Dels	101.7 | 133.7	0.85	0.46
Rare	Dups	242.2 | 57.56	**0.031**	0.15

Table 4.3. 22q11.2 analysis of PennCNV GWAS calls. Total CNV event distance spanned per subject in kb. Significance values of less than 0.05 are highlighted in bold.

Finally, table 4.4 illustrates the average event size per subject. Again, a significant effect for duplications, and principally rare duplications is seen and an absence of effect for deletions. Interestingly, controls have a borderline significant higher average event size than cases for rare deletions. This may not be of true significance as the number of actual events is very small.

Event Frequency	Event Type	KbAvg (cases | controls)	1 sided p value	2 sided p value
All	Dels/Dups	86.03 | 75.32	0.20	0.40
All	Dels	79.09 | 117.5	0.94	0.078
All	Dups	87.51 | 61.23	**0.011**	0.064
Common	Dels/Dups	78.91 | 73.48	0.34	0.63
Common	Dels	79.75 | 110.9	0.92	0.15
Common	Dups	80.11 | 61.55	0.05	0.15
Rare	Dels/Dups	119.6 | 84.08	0.30	0.53
Rare	Dels	66.39 | 133.7	0.94	0.056
Rare	Dups	172.9 | 57.56	**0.045**	0.20

Table 4.4. 22q11.2 analysis of PennCNV GWAS calls. Average CNV event size in kb per subject (KbAvg). Significance values of less than 0.05 are highlighted in bold.

Whilst this analysis supports our original hypothesis in some metrics, it may be confounded by false positive calls that are small, and made with less markers. We decided to validate a subsection of the calls made by PennCNV using our array CGH data.

4.4.2 Follow Up of 22q11.2 PennCNV Calls with array CGH

SUMMARY: Visually validating a subset of 207 randomly selected PennCNV calls with array CGH data, we found that 73.4% were true, and those that did not validate were much more likely to be small, and made with few markers.

We followed up a total of 207 randomly selected calls made by PennCNV in the 22q11.2 region with our array CGH data, visually screening for each CNV manually using OGT's Cytosure software. Of 207 calls, 152 (73.4%) were verified as true, 53 (25.6%) were false positive and 2 (1.0%) could not be verified due to lack of marker coverage on the CGH array. A full table of the CNVs followed up can be found in the Appendix.

We first characterised our followed-up calls by plotting a histogram of call length stratified by copy number (i.e. deletion of duplication) (Fig. 4.2).

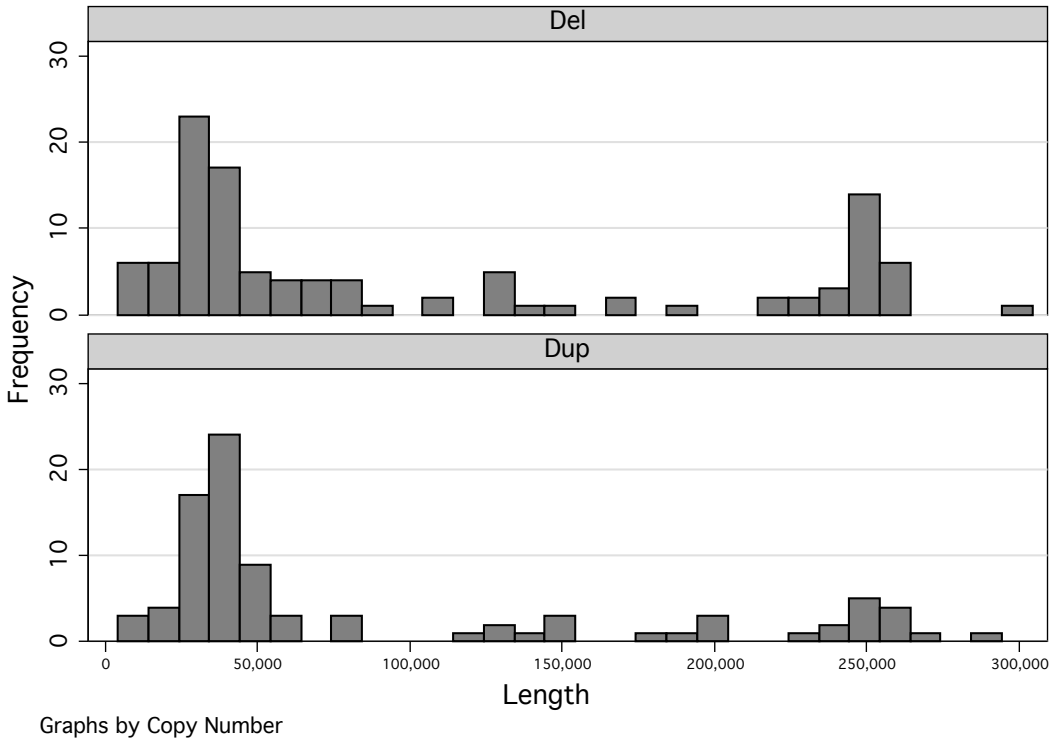


Fig. 4.2. Histograms of CNV calls ordered by size (length in base pairs), stratified into deletions (top panel) and duplications (lower panel).

As can be seen in Fig. 4.3, there is an apparently similar non-parametric distribution of calls. To analyse this statistically we performed a two-sample Wilcoxon rank-sum (Mann-Whitney) test. The two CNV types (deletions and duplications) are no more likely to be different in length than chance ($z=0.393$, $p>|z|=0.39$).

We were next interested to investigate the nature of those calls that were verified compared to those that were not. In absolute terms, of 113 deletions, 74 (65.5%) validated whilst 39 (34.5%) did not, and, of 94 duplications, 78 (83.0%) validated whilst 14 (14.9%) did not and 2 (2.1%) could not be validated.

We next plotted histograms of calls, stratified by validation status, arranged by the number of markers (Fig. 4.3) and length in BP (Fig. 4.4).

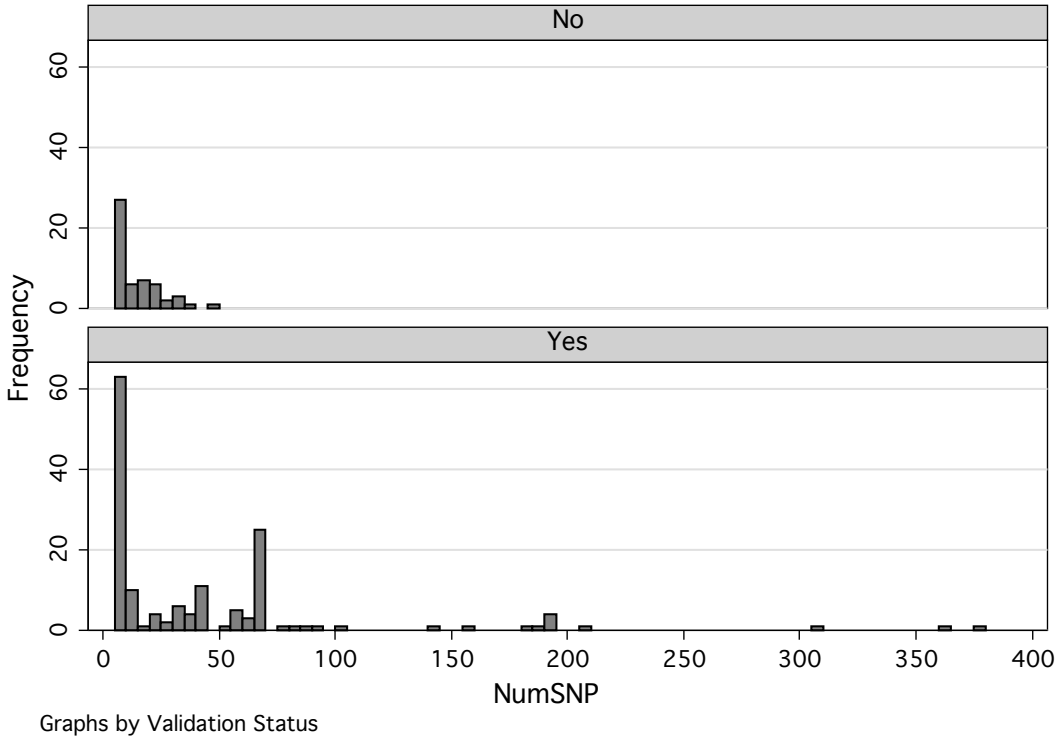


Fig. 4.3. Histograms of CNV calls ordered by the number of markers (NumSNP) used to make them, stratified into those that were validated (lower panel) and those that were not (top panel).

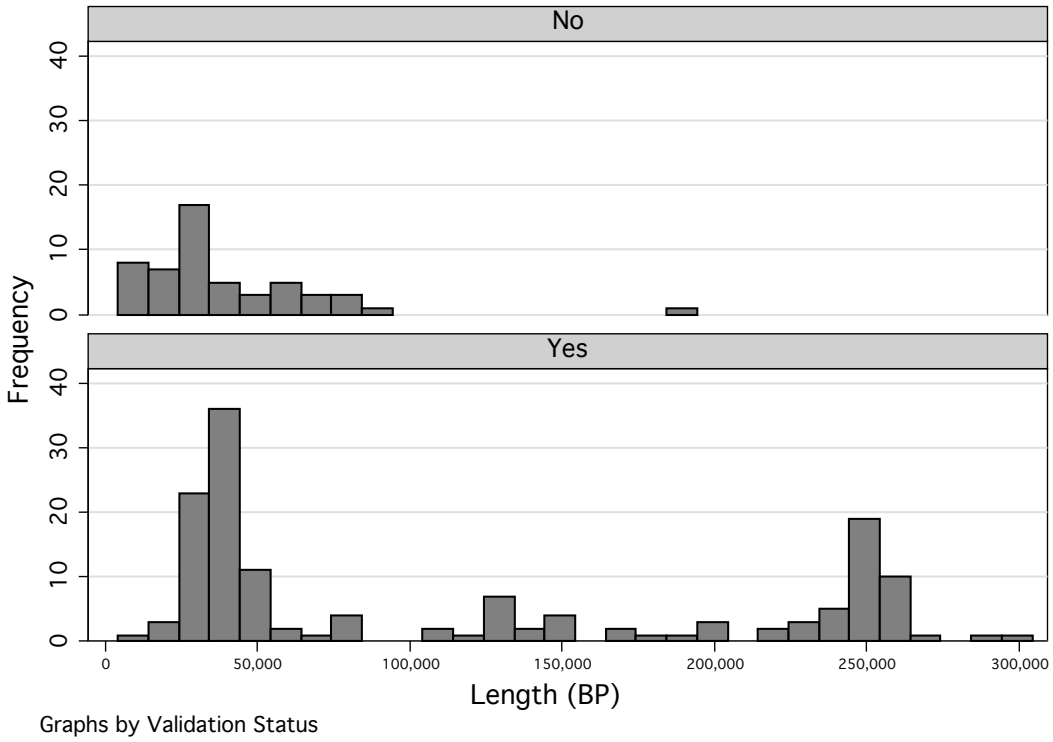


Fig. 4.4. Histograms of CNV calls ordered by size (length in base pairs) and stratified into those that were validated (lower panel) and those that were not (top panel). In this histogram, 6 calls larger than 500,000BP have been omitted (all of which validate) to improve the resolution of the plot.

The histograms indicate that calls that are not validated are more likely to be shorter (Mann-Whitney: $z=-5.65$, $p>|z|<0.0001$) and made with less markers ($z=-3.77$, $p>|z|<0.0002$) than those that were validated. This is intuitively logical and reinforces the QC thresholds imposed on our call set in chapter 2. However this also indicates that differences in QC metrics in some samples may bias analyses that involve smaller calls and do not take account of the validation rate. Hence we decided to analyse our array CGH calls directly to look for significant differences in call rates between cases and controls, similar to our analysis in 4.4.1.

4.4.3 Array CGH Data

SUMMARY: We did not observe significant differences between our cohorts, although this may be because of a lack of statistical power to detect an effect. Given that population stratification has not been accounted for, the results for common CNVs should be interpreted with caution.

Using our array CGH data directly to make copy number calls, we were able to analyse these calls using PLINK for significant association in 4 metrics.

Table 4.5 illustrates the event rate per person, where events denote CNV events. A significant enrichment is seen in common duplications when controls are compared to cases, however as population stratification has not been accounted for in this sample this is not likely to be of relevance.

Event Frequency	Event Type	Rate (cases | controls)	1 sided p value	2 sided p value
All	**Dels & Dups**	3.13 | 3.28	0.70	0.69
All	**Dels**	2.12 | 2.06	0.43	0.81
All	**Dups**	1.00 | 1.22	0.84	0.34
Common	**Dels & Dups**	2.84 | 3.16	0.90	0.22
Common	**Dels**	1.89 | 1.96	0.69	0.71
Common	**Dups**	0.84 | 1.14	0.99	**0.03**
Rare	**Dels & Dups**	0.39 | 0.18	0.28	0.47
Rare	**Dels**	0.24 | 0.098	0.25	0.44
Rare	**Dups**	0.17 | 0.078	0.48	0.65

Table 4.5. Event rate per person (Rate). Significance values of less than 0.05 are highlighted in bold.

Table 4.6 illustrates the proportion of cases and controls to have at least one event (this is similar in nature to our analysis in chapter 2). Logically all samples will contain a CNV event of one type or the other, because the sample set was

selected on that basis. There are, however, no significant differences between the proportion of samples with a deletion or duplication.

Event Frequency	Event Type	Prop (cases | controls)	1 sided p value	2 sided p value
All	**Dels & Dups**	1.00 | 1.00	1.00	1.00
All	**Dels**	0.95 | 0.94	0.56	1.00
All	**Dups**	0.62 | 0.75	0.97	0.11
Common	**Dels & Dups**	0.99 | 1.00	1.00	1.00
Common	**Dels**	0.93 | 0.94	0.67	1.00
Common	**Dups**	0.60 | 0.73	0.97	0.11
Rare	**Dels & Dups**	0.14 | 0.14	0.62	1.00
Rare	**Dels**	0.088 | 0.078	0.54	1.00
Rare	**Dups**	0.062 | 0.058	0.62	1.00

Table 4.6. Proportion of cases/controls to have at least one event (Prop). Significance values of less than 0.05 are highlighted in bold.

Table 4.7 illustrates figures for the total event size per subject. There are no significant differences in any comparison. The absolute difference between the total event size per subject of rare events is strikingly high.

Event Frequency	Event Type	KbTot (cases | controls)	1 sided p value	2 sided p value
All	**Dels & Dups**	56.59 | 38.18	0.17	0.35
All	**Dels**	42.08 | 24.90	0.17	0.35
All	**Dups**	26.94 | 19.78	0.44	0.72
Common	**Dels & Dups**	38.98 | 37.85	0.47	0.92
Common	**Dels**	29.01 | 19.25	0.18	0.37
Common	**Dups**	14.86 | 20.04	0.92	0.16
Rare	**Dels & Dups**	156.7 | 40.23	0.19	0.35
Rare	**Dels**	151.0 | 67.81	0.44	0.69
Rare	**Dups**	128.2 | 3.46	0.097	0.28

Table 4.7. Total event distance spanned per subject (KbTot). Significance values of less than 0.05 are highlighted in bold.

Finally, table 4.8 illustrates the average event size per subject. Again, there are no significant differences in any comparison, although, again the absolute difference in the average rare event size per subject is strikingly high, although in this analysis only for rare duplications.

Event Frequency	Event Type	KbAvg (cases | controls)	1 sided p value	2 sided p value
All	**Dels & Dups**	16.82 | 9.73	0.12	0.28
All	**Dels**	13.29 | 7.89	0.15	0.29
All	**Dups**	16.00 | 13.05	0.60	0.77
Common	**Dels & Dups**	12.22 | 10.21	0.32	0.59
Common	**Dels**	11.46 | 7.25	0.19	0.40
Common	**Dups**	9.88 | 13.20	0.89	0.20
Rare	**Dels & Dups**	73.11 | 21.90	0.28	0.50
Rare	**Dels**	45.55 | 35.86	0.55	0.87
Rare	**Dups**	96.60 | 3.28	0.13	0.34

Table 4.8. Average event size per subject (KbAvg). Significance values of less than 0.05 are highlighted in bold.

Overall, these results are disappointing in that they do not show the expected association between rare deletions and caseness. However given that our aCGH sample is well over 10 fold smaller than our original analysis, we may lack the power to detect an effect if one exists.

4.4.4 Delineation of the 22q11.2 Deletion

SUMMARY: aCGH data validates the large deletion CNV seen in 22q11.2 as a typical 3MB deletion seen in a variety of clinical syndromes, including Velocardiofacial syndrome.

The most interesting CNV in our dataset is a large 22q11.2 deletion called across a 3MB region seen in a variety of eponymous syndromes, most notably Velocardiofacial syndrome (VCFS), DiGeorge syndrome (DGS) and conotruncal anomaly facial syndrome (CAFS) (Robin & Shprintzen, 2005). Fig. 4.5 illustrates the call made by PennCNV in our Illumina 610 Quad chip data and the corresponding call in our aCGH data.

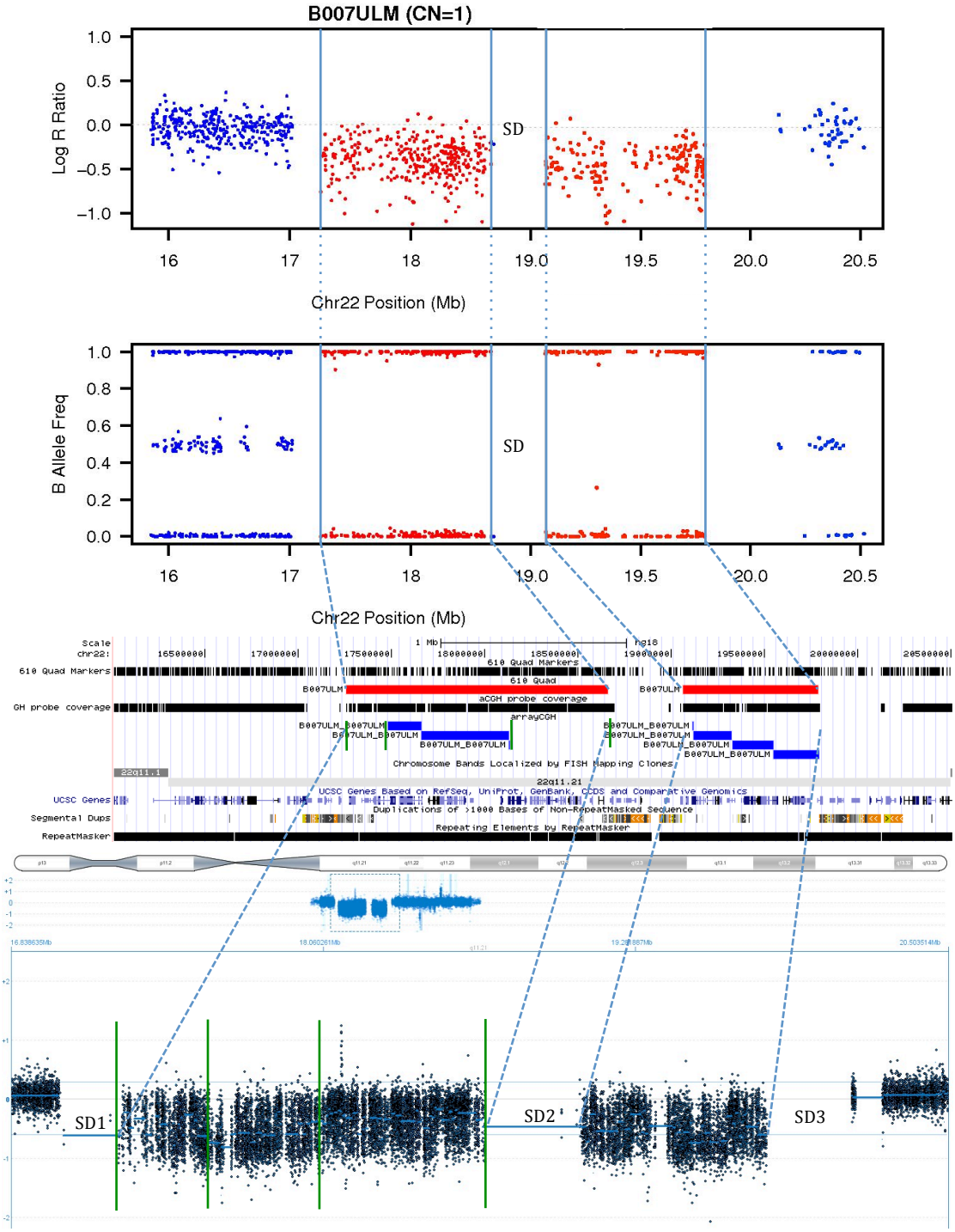


Fig. 4.5. A large deletion CNV in chromosome 22q11.2 is called around three areas of segmental duplication (SD1-3). Array CGH data (bottom panel) supports the presence of one long CNV divided by the central area of segmental duplication, however the DNAcopy algorithm calls this CNV as different segments, and misses a large proximal and central (green lines indicate boundaries in bottom and central panel) portion of the CNV (compare blue line and red line, central panel).

Interestingly, this call is not completely supported by our DNAcopy calls when viewed on the UCSC browser as a custom BED file track (Fig. 4.6, central panel, blue lines). However on visual inspection of the array CGH data (Fig. 4.6, bottom panel) the CNV clearly runs for the full length predicted by PennCNV. This is an interesting observation, which may happen for a number of reasons. Firstly it highlights the problems of calling CNVs with noisy fluorescence data using segmentation methods, and secondly it highlights the difficulties in applying thresholds with which to call regions of differing copy number. In our analysis we assumed a normal distribution of segments after excluding the majority of segments we reasonably thought to represent normal copy number, and then set a threshold at 2 x SD less than the mean of the remaining segments. This was based on the observation that setting the threshold at 3 x SD included a proportion of segments with normal copy number. This, in itself, again highlights the essential problem in CNV analysis- the balance between making type 1 (false positive) and type 2 (false negative) errors. In fact the array CGH data does suggest a slight deviation in the mean relative probe intensity between 18MB and 18.75MB in Fig. 4.6. In absolute terms, the mean segment probe intensity for the un-called deletion region is -0.1017, whilst the mean segment probe intensity for the called deletion region is -0.1707. This observation may be reconciled by misbehaving probes in the affected area. This area of chromosome 22 has been evolutionarily subject to duplication, so probes in this area may also partially hybridise DNA from other areas of the genome. It could also be possible that this sample is mosaic for the deletion at this un-called region, which is also suggested by a very slight departure of the values from 0 and 1 in the B allele frequency

plot in the top panel of Fig. 4.5. If this sample is mosaic, it may explain the relatively mild phenotype in comparison to other subjects with 22q11.2 deletions of this size, which will be illustrated more extensively in chapter 6.

The array CGH data, on visual inspection shows that the CNV in question is called with breakpoints at precisely the regions one would expect given the boundaries of the flanking regions of segmental duplication. The proximal breakpoint called with aCGH is chr22:17,255,830bp, whilst the distal breakpoint is called at chr22:19,794,056bp. This compares to a proximal breakpoint called with the Illumina 610 Quad data of chr22:17,257,786bp and a distal breakpoint of chr22:19,792,353bp. Thus the Illumina array called CNV is highly similar to that called by array CGH.

The proximal break point does not interrupt any known gene coding region. The distal breakpoint interrupts a putative breakpoint cluster region (BCR) like protein (Fig. 4.6), which is similar to the BCR gene also on chromosome 22 that has one reported SNP association with affective disorders in a candidate gene study (Hashimoto et al., 2005). However it seems unlikely that this is a major contributory factor in this case.

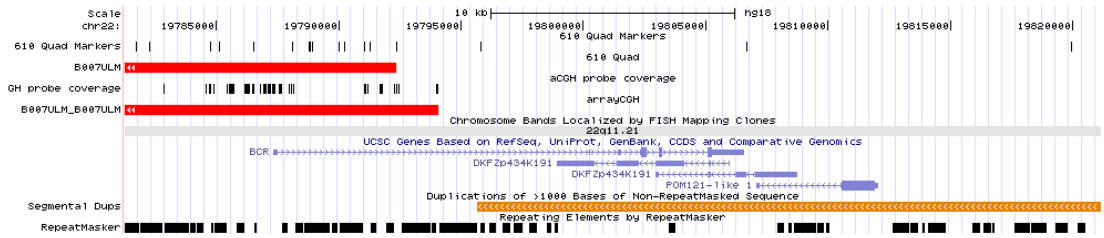


Fig. 4.6. The distal breakpoint of the large 22q11.2 deletion CNV interrupts a putative breakpoint cluster region like protein.

4.5 Conclusion

We hypothesised A) that our association between rare deletions and caseness would be retained in a re-analysis of our 22q11.2 data. This was partially supported by our data, however on follow up of a subset of our calls with our array CGH data we were able to show that smaller CNV calls, and calls made with low numbers of markers, are less likely to be validated than larger calls.

We hypothesised B) that our findings would be supported by calls derived from a randomly selected subset of samples followed on the high density CGH array. We were not able to replicate this result with our array CGH data, perhaps because of inadequate power to detect an effect.

Finally we characterised the nature of the 22q11.2 deletion seen in 1 of our cases, and found that it is typical of the large deletion that is often seen in cases of velocardiofacial syndrome.

Chapter 5. iPattern and QuantiSNP Calling Methods

5.1 Introduction

PennCNV is an attractive, easy-to-use package for CNV calling, however alternative methods for calling CNVs are available. Studies of CNV call reproducibility across methods suggest that the concordance between methods is quite low, although any quoted figure varies considerably as a function of call size and number of markers upon which a call is based(Pinto et al., 2011; D. W. Tsuang et al., 2010; Winchester, Yau, & Ragoussis, 2009). In this final analysis chapter I describe the approaches implemented by the QuantiSNP and iPattern methods. I then go on to describe results of rare CNV analyses with these methods, the effect of intersecting calls from three methods to reduce false positive call rates and the effects of using further quality control metrics other than those already described.

5.1.3 CNV Call Reproducibility

It is worth pointing out at this stage that research into the reproducibility of copy number calls has generally shown that different methods can produce quite different results(Dellinger et al., 2010; Eckel-Passow, Atkinson, Maharjan, Kardia, & de Andrade, 2011; D. W. Tsuang et al., 2010; Winchester et al., 2009). Such research tends to be conflicting with different groups recommending different methods in different sets of circumstances, and settings within different methods not being consistent between studies. Some have argued for agreed standards in studies of structural variation in the human genome(Scherer et al., 2007). In the face of rapidly changing markets and competing technologies, this so far has not been forthcoming.

In the absence of a consensus methodology on CNV calling, and given the above concerns, it is logical to expect that the false positive rate for CNV calls will be reduced if a call set is restricted to overlapping calls in the same sample made with more than one method. In collaboration with the Scherer laboratory in Toronto, we decided to do this using consensus call boundaries from the QuantiSNP and iPattern methods, validated with PennCNV.

5.1.2 QuantiSNP

QuantiSNP is developed by Christopher Yau and Ioannis Ragoussis in Oxford(Colella et al., 2007). In contrast to PennCNV, QuantiSNP uses an objective Bayesian method to set certain prior parameters for the calling model and calculate probabilities of each copy number state at each marker, although both use a hidden-Markov model to call CNVs, and both have been developed to work with Illumina BeadArray data. An objective Bayesian method has some advantages when one of the hidden states, in this case normal copy number, or the null state, can be assumed to be occurring much of the time. Within QuantiSNP a pre-existing parameter is set which dictates the degree of expectation that the model will deviate from the state of normal copy number. This may reduce the frequency of type 1 errors, although in consequence may also increase the rate of type 2 errors.

5.1.3 iPattern

We also processed our samples using iPattern, a method for CNV calling developed by Dalila Pinto and Zhuozhi Wang at the Centre for Applied Genomics in Toronto. iPattern uses non-parametric density-based clustering to categorize

sample loci into different groups when compared to reference samples by using an moving window based approach. The largest cluster of unrelated samples is chosen as a reference, and samples with higher or lower intensities are assigned as relative CNV gains or losses (unpublished). In this sense the method is different to both the PennCNV and QuantiSNP methods, because it does not rely on Illumina's GenTrain algorithm for clustering markers and deriving relative and allelic intensity ratios. Whilst iPattern has been used in some high-profile CNV studies(Fernandez et al., 2010; Pinto et al., 2010) the methodology remains unpublished and cannot yet be said to be well-established in the research community.

5.2 Hypotheses

We hypothesised that

A) The association between cases and rare deletion CNVs would be retained by using data derived from QuantiSNP, iPattern and a set of consensus calls made from intersecting calls from all three methods.

5.3 Methods

Samples and genotyping methods have already been described in chapter 2. In our genome wide burden analyses we continued to restrict our sample set to those derived from venous blood and cell lines, excluding cheek swab calls, as in chapter 2. However in a later analysis of regions previously implicated in

schizophrenia, presented in 5.4.4, we included calls from cheek swab samples because it was practical to visually follow up each call.

5.3.1 CNV Calling

5.3.1.1 QuantiSNP

We processed our GenomeStudio-derived LRR and BAF values with QuantiSNP (v2.3) using the settings suggested by the authors and also after liaison with Christopher Yau to check that this was suitable for our dataset . These settings comprise a list of parameters, which may be viewed in the appendix. Because our data was based on Illumina infinium data, which is similar to data used to model the prior parameters for the QuantiSNP method, we did not have to alter these settings. QuantiSNP calculates a maximum log Bayes factor as a calculation of the confidence of the CNV call. We used a threshold of 15 for our calls, as recommended by the author. Unlike PennCNV, QuantiSNP does not process samples in serial and we therefore wrote a Perl script to run each sample separately on a high performance linux cluster (see appendix). This was advantageous as the QuantiSNP method is otherwise very slow, taking over 3 months to process our samples on a standard desktop workstation. Using processing of samples in parallel we were able to considerably shorten this time.

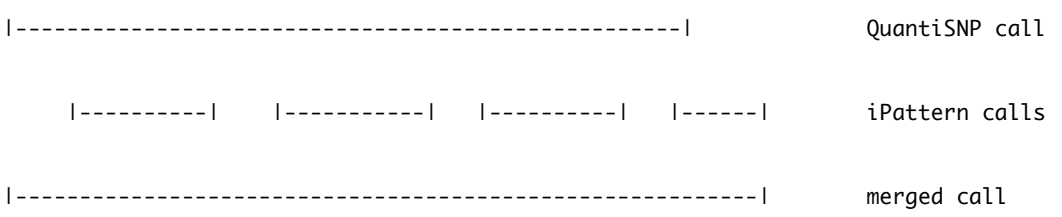
5.3.1.2 iPattern

The iPattern method has not yet been published in a peer-reviewed journal, although has been used in several high profile CNV studies(Korn et al., 2008; Pinto et al., 2011). We processed normalised X and Y values (representing within sample normalised fluorescence values) derived from GenomeStudio with

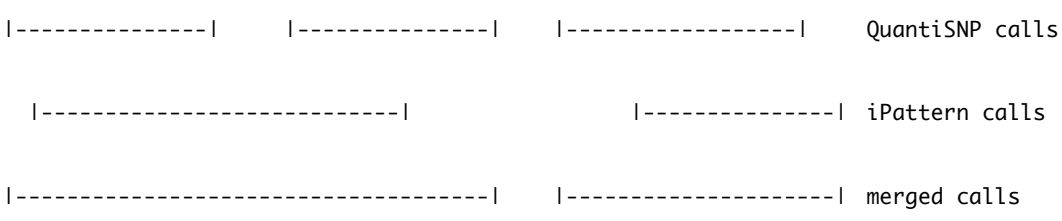
iPattern. iPattern takes X and Y values and uses a pre-processing method to evaluate the background signal to noise ratio for each batch of tested samples. Outliers from the standard deviation of the sample batch are removed. A two-stage analytical framework is then used to identify CNV regions, with a moving window-based approach followed by secondary boundary refinement. The largest cluster of unrelated samples is dynamically chosen as reference, and samples with higher or lower intensities are assigned as relative CNV gains or losses. CNV lengths are calculated based on the distance between the first and last array probes internal to the variant(Pinto et al., 2010). Copy number calls were made with a minimum of 5 markers and denoted to be either a 'loss' or a 'gain' depending on the spread of the B Allele Frequency markers. Of note, iPattern does not distinguish between homozygous and hemizygous CNV calls.

5.3.1.3 Intersection

We derived a call set based on the intersection of the outside boundaries of overlapping calls from the same sample made by iPattern and QuantiSNP, with PennCNV calls used for confirmation of calls, but not for boundary delineation. Hence all calls used in this analysis were restricted to those made by all three methods, with the exception of singleton calls (that is, calls made only once throughout a single methods calls) which were made if they were called with either iPattern or QuantiSNP and overlapped with a PennCNV call made with the same sample. Calls made as a continuous event by one method but which were broken by another method were merged into a consensus call, as illustrated



However if the calls were fragmented by both programs, such regions were not merged.



We used this call set in a further analysis of our data.

5.3.1.4 Further Sample Quality Control

We used quality control (QC) thresholds already set with our PennCNV data in order to derive a comparable sample set and call set (see 2.3.5). We also performed a high QC analysis, in which we used a number of other metrics to further refine our sample set and exclude samples that may be contributing false positive calls. We used metrics and set thresholds as follows,

1. Penn WF (waviness factor) <0.04 and >-0.04

2. Extreme LRR values > 5,627 (>1% of total number of markers)

3. Wide LRR values > 28,135 (>5% of total number of markers)

4. Number of CNV calls (before any QC) made by any method > 3SD from the mean call number.

The waviness factor is a metric calculated by PennCNV based on the median absolute deviation of LRR values, and designed to be robust to outliers (as, for example, may occur with the occurrence of a true CNV)(Diskin et al., 2008). Extreme and wide LRR values are metrics designed to detect samples with excessive numbers of extreme (<-1) and wide (deviation from 0 of >0.5) LRR values respectively. Array problems, such as window mismatching during fluorescence reading, can generate samples with high genotype call rates and relatively good basic QC (LRRSD and BAFSD) metrics, but on visual inspection the samples have excessive extreme (Fig. 5.1) and wide (Fig. 5.2) LRR values. We calculated extreme and wide LRR values using scripts extracted from the CNVision package (http://futo.cs.yale.edu/mw/index.php/CNVision), and using the limits above as advised by the Scherer lab. Finally we excluded any sample where the number of calls made by any method exceeded 3xSD from the mean number of calls across all samples, before any call QC. In practice this meant excluding any sample with more than 153 PennCNV calls, 81 QuantiSNP calls and 110 iPattern calls.

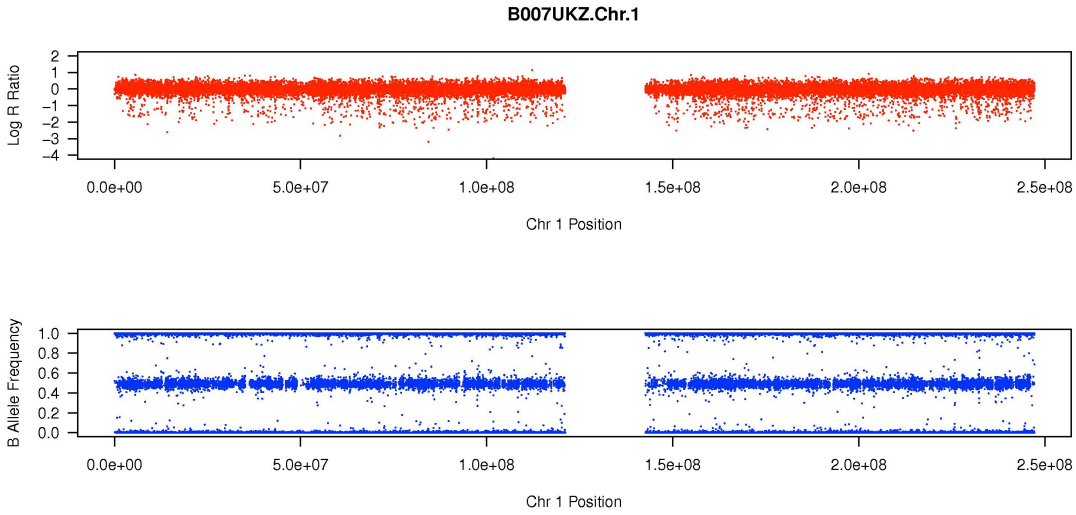


Fig. 5.1. LRR and BAF plot of sample B007UKZ, chromosome 1, with a genotype call rate of 99.7%, a LRRSD of 0.257 and a BAFSD of 0.030, has a significant number (9,319) of extreme (<-1) LRR values (upper plot) and is excluded in our high QC analyses.

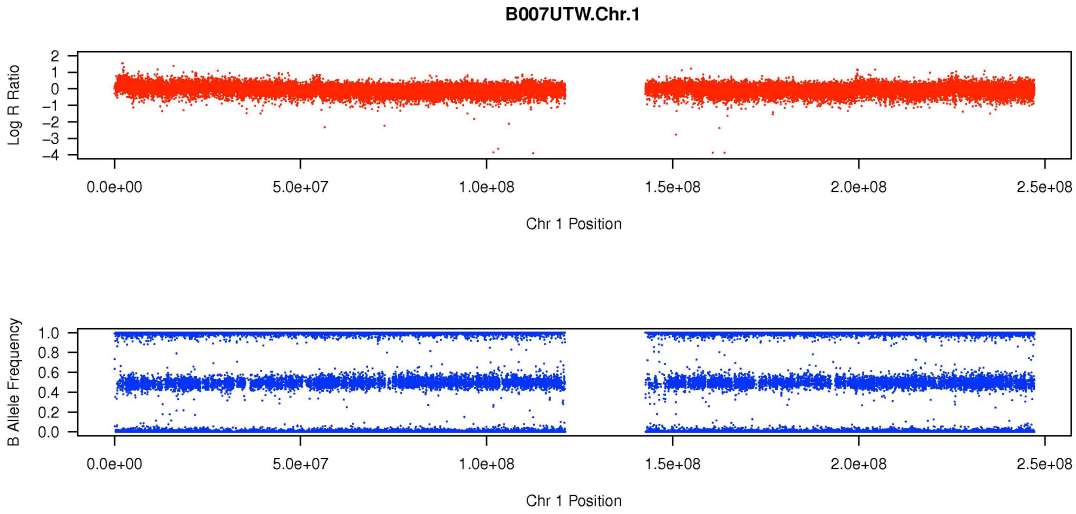


Fig. 5.2. LRR and BAF plot of sample B007UTW, with a genotype call rate of 99.9%, a LRRSD of 0.269 and a BAFSD of 0.0386, has a significant number (36,747) of wide (>0.5 or <-0.5) LRR values (in this case due to waviness) and is excluded from our high QC analyses.

Finally, we analysed data across samples by chromosome to exclude samples with total and mosaic aneuploidy, which may, for example, be introduced in cell line creation or be a function of age. 7 samples, all from the 1958 birth cohort,

and all containing duplication aneuploidies in autosomes were removed from our analysis.

5.3.2 CNV Analysis

In all of our analyses in this chapter we used PLINK(Purcell et al., 2007) with the --mperm and --cnv-test-2sided functions to calculate empirical 1 and 2 sided p values with 10,000 null permutations of case-control status. This is different to our original analysis method in chapter 2, where we calculated Fisher's exact test for the number of samples with different types of CNVs occurring over various regions of the genome. PLINK provides a fast, convenient and statistically robust method for calculating numerous CNV metrics throughout a sample set.

We also used PLINK to exclude common CNVs in this analysis, as PLINK provides a more robust method to identify regions of copy number polymorphism with calls that only partially overlap. Because of this, burden calculations for the case sample vary slightly as differing CNP regions are excluded with call sets from different methods based on the 1% threshold. Put another way, because CNV calls vary between methods, so will the CNP regions excluded by PLINK vary between methods. We present results for all methods used, including PennCNV, with the observation that the frequencies calculated by PLINK for PennCNV data will vary slightly from our results in chapter 2. Otherwise, our call sets were processed in an identical fashion to that described in 2.3.5.

PLINK does not provide a function for calculating p-values for CNVs in purely exonic regions, however we used the --cnv-count function to calculate the number of genes spanned by CNVs, the number of CNVs with at least one gene

and the number of genes per kilobase of CNV, along with empirical 1 sided and 2 sided p-values.

5.4 Results

5.4.1 Results from PennCNV, iPattern and QuantiSNP Call Sets

The range of numbers of calls made by the three methods across samples can be seen in table 5.1, and is visually represented in Figs. 5.3, 5.4 and 5.5. To improve the resolution of the histograms, samples with more than 100 calls have been removed from the figures.

Method	Mean no. of Calls	Standard Deviation	Range
PennCNV	21.2	42.7	1 - 1,348
iPattern	13.0	31.7	1 - 496
QuantiSNP	7.7	23.5	0 - 2,172

Table 5.1. Summary statistics for the number of calls made by each method, across all cohorts. NB Calls made with at least 5 markers. Cheek swab samples excluded.

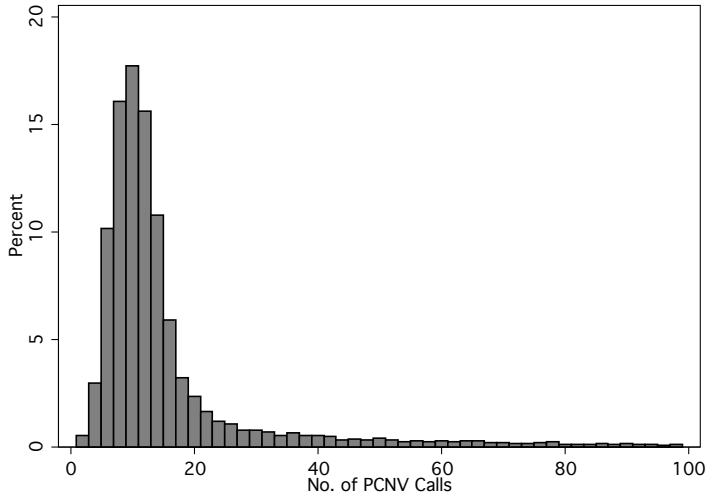


Fig. 5.3. Histogram of the number of PennCNV calls made across all samples.

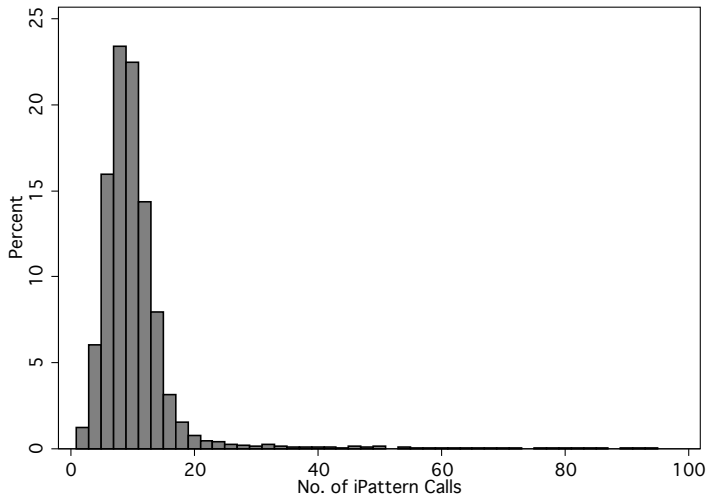


Fig. 5.4. Histogram of the number of iPattern calls made across all samples.

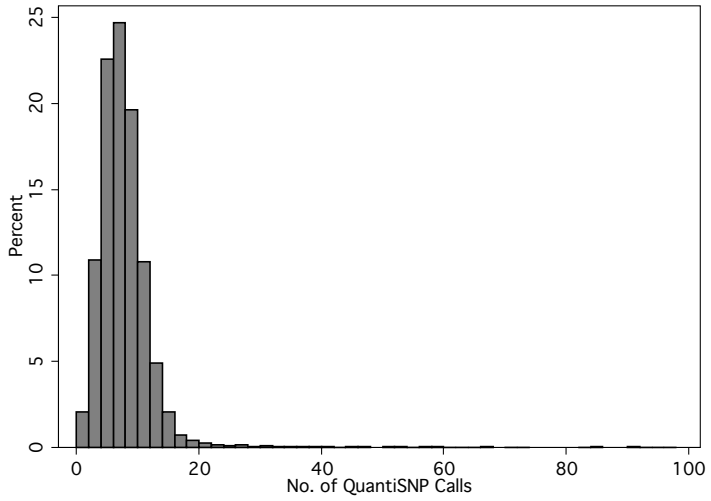


Fig. 5.5. Histogram of the number of QuantiSNP calls made across all samples.

The range of average call length made by the three methods across samples is presented in table 5.2, and is visually represented in Figs. 5.6, 5.7 and 5.8. To improve the resolution of the histograms, samples with more than 10MB of CNV calls have been removed from the figures.

Method	Mean Length of Calls (Kb)	Standard Deviation (Kb)	Range (Kb)
PennCNV	2,050	6,911	9,200 - 30,700
iPattern	1,151	3,050	4,600 - 102,000
QuantiSNP	1,409	33,100	0 - 2,800,000

Table 5.2. Summary statistics for the number of calls made by each method. NB Calls made with at least 5 markers. Cheek swab samples excluded.

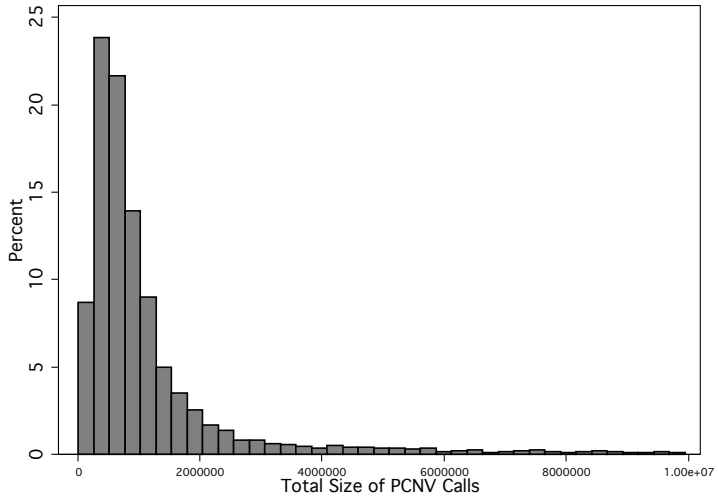


Fig. 5.6. Histogram of the total size of CNV calls, in bp, per sample across cohorts made by the PennCNV methods

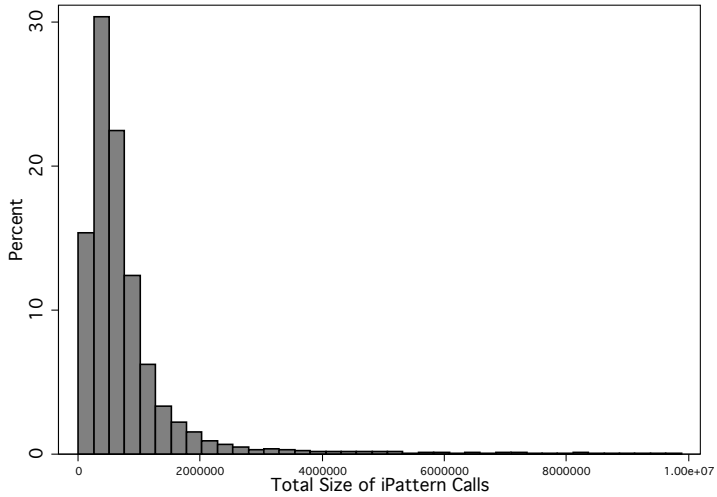


Fig. 5.7. Histogram of the total size of CNV calls, in bp, per sample across cohorts made by the iPattern method

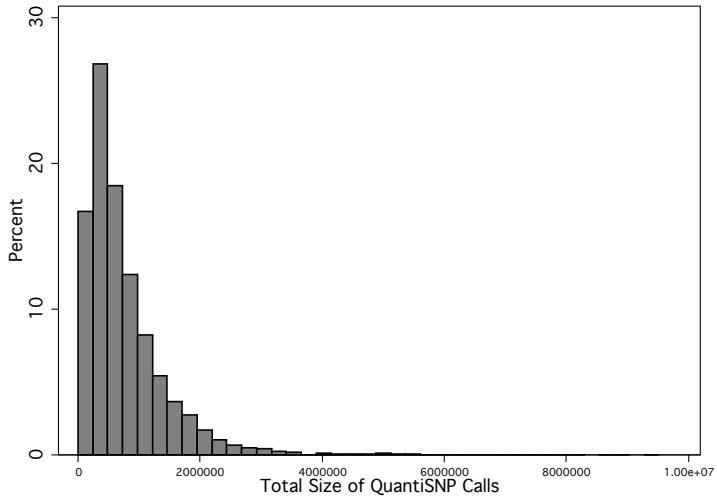


Fig. 5.8. Histogram of the total size of CNV calls, in bp, per sample across cohorts made by the QuantiSNP method

Initially we processed our data from each method separately using PLINK, and compared the results for each of the seven metrics reported by PLINK-

1. RATE- The number of CNVs per sample

2. PROP- The proportion of samples with at least 1 CNV

3. TOTKB- The total length of CNVs in kilobases (kb).

4. AVGKB- The average length of CNVs in kilobases.

5. GRATE- Number of genes spanned by CNVs.

6. GPROP- Number of CNVs with at least one gene.

7. GRICH- Number of genes per kilobase of CNV.

After QCing our calls we analysed our cohorts using both the standard QC sample set described in 2.3.4 and the high QC sample set described in 5.3.1.4.

5.4.1.1 Standard QC Threshold Results

5.4.1.1.1 Cases Vs. Screened Controls

SUMMARY: Within the analysis of data from each method separately, there is additional support from the iPattern method for the original notion from chapter 2 that rare deletion CNVs are significantly more frequent in cases when compared both to screened controls and WTCCC2 control. However this is not supported by the QuantiSNP method

2,723 cases and 348 screened controls passed our standard QC thresholds and were used in this analysis.

We initially compared the results of each method in our cases and screened controls.

Table 5.3 illustrates the CNV rate per sample. Both PennCNV and iPattern call a significantly higher number of CNVs in cases than controls ($p<0.0001$ & $p=0.0086$ respectively), with a particular enrichment for deletions ($p<0.0001$ & $p=0.0091$ respectively). There is no significant difference when calls from the QuantiSNP method are analysed ($p=0.65$ for all CNVs & $p=0.68$ for deletion CNVs), and this method calls considerably less CNVs than either iPattern or PennCNV.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	1.32 | 0.79	**0.0001**	**0.0001**	0.95 | 0.76	**0.0086**	**0.021**	0.78 | 0.80	0.65	0.77
Rare Del	0.92 | 0.44	**0.0001**	**0.0001**	0.53 | 0.38	**0.0091**	**0.029**	0.34 | 0.35	0.68	0.71
Rare Dup	0.52 | 0.45	0.090	0.18	0.57 | 0.45	**0.041**	0.10	0.48 | 0.48	0.56	0.95

Table 5.3. Standard QC. Cases vs. screened controls. Standard QC. Cases vs. screened controls. CNV event rate (Rate) per person. Significance values of less than 0.05 are highlighted in bold.

Table 5.4 illustrates the proportion of cases and controls to have at least 1 CNV event. In this analysis significantly more cases than controls have a rare deletion CNV when calls from PennCNV are used (p=0.01), however this is not supported by the iPattern and QuantiSNP methods. The degree of association with PennCNV is reduced from the analysis presented in 2.4.1 (p=0.0005 vs. p=0.01) which is driven by a slightly reduced number of cases with deletions and a slightly greater number of screened controls with deletions (0.41 vs. 0.40 for cases & 0.32 vs. 0.34 for screened controls). This may reflect either or both of the method used to identify areas of copy number polymorphism, and the statistical method implemented in PLINK to analyse for association.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.58 | 0.52	**0.025**	0.053	0.52 | 0.51	0.46	0.91	0.52 | 0.53	0.67	0.74
Rare Del	0.40 | 0.34	**0.01**	**0.017**	0.32 | 0.31	0.36	0.67	0.28 | 0.30	0.77	0.53
Rare Dup	0.35 | 0.34	0.46	0.86	0.32 | 0.34	0.75	0.55	0.34 | 0.34	0.55	1.00

Table 5.4. Standard QC. Cases vs. screened controls. Proportion of cases/controls to have at least one CNV event (Prop). Significance values of less than 0.05 are highlighted in bold.

Table 5.5 illustrates the total CNV event distance per subject. Here, both PennCNV and iPattern detect a highly significant difference between cases and screened controls (p=0.0003 & p=0.0018 respectively), principally driven by rare deletions (p=0.0001 & p=0.0008 respectively). The QuantiSNP method shows a non-significant trend in the same direction (p=0.092).

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	524 | 372	**0.0003**	**0.0016**	498 | 367	**0.0018**	**0.014**	418 | 378	0.16	0.33
Rare Del	456 | 268	**0.0001**	**0.0017**	436 | 266	**0.0008**	**0.017**	295 | 243	0.092	0.23
Rare Dup	417 | 356	0.072	0.15	498 | 357	**0.0042**	**0.025**	417 | 387	0.27	0.51

Table 5.5. Standard QC. Cases vs. screened controls. Total CNV event distance spanned per subject in kb (KbTot). Significance values of less than 0.05 are highlighted in bold.

Table 5.6 illustrates the average CNV event size per subject. Here, none of the three methods detect a significant difference between cases and controls, suggesting that, if they do occur, rare CNVs are no more likely to be of different size in cases or controls.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	247 | 249	0.59	0.90	262 | 256	0.40	0.74	267 | 251	0.23	0.44
Rare Del	213 | 211	0.52	0.95	250 | 218	0.11	0.23	233 | 203	0.13	0.28
Rare Dup	283 | 274	0.36	0.68	279 | 281	0.54	0.96	290 | 281	0.37	0.7

Table 5.6. Standard QC. Cases vs. screened controls. Average CNV event size in kb per subject (KbAvg). Significance values of less than 0.05 are highlighted in bold.

Table 5.7 illustrates the number of genes spanned by CNVs. Again, the PennCNV method suggests that case subjects are significantly more likely to have a deletion CNV spanning a gene than controls (p=0.0006), principally driven by deletions (p=0.0003). The degree of this association is reduced in the iPattern method (p=0.029 for all CNVs, p=0.08 for deletions and p=0.052 for duplications). The QuantiSNP method does not statistical support this association (p=0.56 for all CNVs).

	PennCNV			**iPattern**			**QuantiSNP**		
Event Frequency & Type	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.32 | 0.19	**0.0006**	**0.0034**	0.19 | 0.14	**0.029**	0.067	0.16 | 0.16	0.56	1.00
Rare Del	0.22 | 0.11	**0.0003**	**0.0041**	0.096 | 0.070	0.08	0.15	0.071 | 0.067	0.43	0.84
Rare Dup	0.10 | 0.081	0.18	0.33	0.11 | 0.070	0.054	0.11	0.094 | 0.096	0.57	1.00

Table 5.7. Standard QC. Cases vs. screened controls. The number of genes spanned by CNV events (GRate). Significance values of less than 0.05 are highlighted in bold.

Table 5.8 illustrates the number of CNVs involving at least one gene. Again, the PennCNV method supports the association of rare deletion CNVs with genes in cases over controls (p=0.0026), but this is not supported by the QuantiSNP method (p=0.51), and the iPattern result fails to reach statistical significance (p=0.13).

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.22 | 0.17	**0.0068**	**0.014**	0.16 | 0.13	0.069	0.14	0.14 | 0.15	0.68	0.75
Rare Del	0.15 | 0.096	**0.0026**	**0.0051**	0.089 | 0.069	0.13	0.23	0.068 | 0.067	0.51	0.91
Rare Dup	0.087 | 0.075	0.27	0.48	0.088 | 0.067	0.11	0.2	0.084 | 0.087	0.60	0.92

Table 5.8. Standard QC. Cases vs. screened controls. The number of CNV events involving at least one gene (GProp). Significance values of less than 0.05 are highlighted in bold.

Finally, table 5.9 illustrates the number of genes within each kilobase of CNV. No method supports a significant association in this metric.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.0014 | 0.0013	0.22	0.42	0.0013 | 0.00096	0.052	0.12	0.0012 | 0.0011	0.47	0.92
Rare Del	0.0016 | 0.0013	0.18	0.35	0.0013 | 0.0011	0.18	0.38	0.0014 | 0.0011	0.16	0.35
Rare Dup	0.0011 | 0.0010	0.39	0.73	0.0011 | 0.00090	0.24	0.48	0.0011 | 0.0011	0.45	0.88

Table 5.9. Standard QC. Cases vs. screened controls. Number of genes per total CNV kb (GRich). Significance values of less than 0.05 are highlighted in bold.

5.4.1.1.2 Cases Vs. WTCCC2 Controls

2,723 cases and 4,828 WTCCC2 controls passed our standard QC thresholds and were included in this analysis.

We then compared the results of each method in our cases and the WTCCC2 controls.

Table 5.10 illustrates the CNV event rate per person. In this analysis both PennCNV and iPattern support the association of rare deletions (p=0.0066 & p=0.0005 respectively), and in the case of iPattern rare duplications (p=0.01) with cases. The QuantiSNP method does not support this, even having a significant association in the other direction (2 sided p=0.015)

	PennCNV			**iPattern**			**QuantiSNP**		
Event Frequency & Type	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	1.35 | 1.22	**0.0096**	**0.017**	0.95 | 0.86	**0.01**	**0.019**	0.77 | 0.77	0.49	0.97
Rare Del	0.94 | 0.82	**0.0066**	**0.0095**	0.58 | 0.47	**0.0005**	**0.0007**	0.33 | 0.37	0.99	**0.015**
Rare Dup	0.54 | 0.53	0.36	0.73	0.57 | 0.50	**0.01**	**0.02**	0.49 | 0.49	0.48	0.97

Table 5.10. Standard QC. Cases vs. WTCCC2 controls. CNV event rate per person (Rate). Significance values of less than 0.05 are highlighted in bold.

Table 5.11 illustrates the proportion of subjects to have at least one CNV event. Again, the PennCNV and iPattern methods support the association between rare deletions and cases (p=0.0004 & p=0.008 respectively), whilst the QuantiSNP method suggests a significant association in the other direction (2 sided p=0.012).

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.58 | 0.55	**0.0025**	**0.0033**	0.51 | 0.49	**0.014**	**0.026**	0.51 | 0.50	0.15	0.30
Rare Del	0.41 | 0.37	**0.0004**	**0.0005**	0.33 | 0.31	**0.008**	**0.017**	0.27 | 0.30	1.00	**0.012**
Rare Dup	0.35 | 0.35	0.33	0.66	0.33 | 0.30	**0.011**	**0.022**	0.35 | 0.34	0.15	0.30

Table 5.11. Standard QC. Cases vs. WTCCC2 controls. Proportion of cases/controls to have at least one CNV event (Prop). Significance values of less than 0.05 are highlighted in bold.

Table 5.12 illustrates the total CNV event distance in kilobases per subject. Interestingly, only the iPattern method supports the association between rare deletions and cases in this instance (p=0.0007). Neither the PennCNV or QuantiSNP methods support this observation.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	530 | 542	0.57	0.77	493 | 448	**0.017**	**0.026**	421 | 502	0.75	0.51
Rare Del	461 | 445	0.25	0.51	467 | 372	**0.0007**	**0.0005**	296 | 291	0.38	0.79
Rare Dup	426 | 466	0.83	0.40	491 | 465	0.17	0.34	419 | 540	0.80	0.50

Table 5.12. Standard QC. Cases vs. WTCCC2 controls. Total CNV event distance spanned per subject in kb (KbTot). Significance values of less than 0.05 are highlighted in bold.

Table 5.13 illustrates the average CNV event size in kilobases per subject. As in our previous analysis with screened controls, there is no difference between cases and WTCCC2 controls via any of the three methods.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	246 | 256	0.91	0.18	260 | 251	0.15	0.30	269 | 273	0.59	0.76
Rare Del	212 | 218	0.76	0.48	253 | 239	0.058	0.11	234 | 233	0.45	0.94
Rare Dup	288 | 299	0.86	0.26	273 | 276	0.59	0.78	290 | 297	0.63	0.69

Table 5.13. Standard QC. Cases vs. WTCCC2 controls. Average CNV event size in kb per subject (KbAvg). Significance values of less than 0.05 are highlighted in bold.

Table 5.14 illustrates the number of genes spanned by CNV events in cases and WTCCC2 controls. PennCNV, and to a lesser extent, iPattern, support the association between number of genes spanned by deletion CNVs (p=0.0007 & p=0.045 respectively) and cases, but this is not supported by the QuantiSNP method (p=0.87). There are significantly more cases with rare duplication events called by the iPattern method (p=0.032), but this is not supported by the PennCNV or QuantiSNP methods (p=0.55 & p=0.84 respectively).

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.32 | 0.28	**0.0067**	**0.014**	0.19 | 0.17	0.097	0.18	0.16 | 0.18	0.93	0.16
Rare Del	0.23 | 0.18	**0.0007**	**0.0014**	0.11 | 0.095	**0.045**	0.085	0.071 | 0.078	0.87	0.30
Rare Dup	0.10 | 0.10	0.55	0.92	0.11 | 0.090	**0.032**	0.061	0.094 | 0.10	0.84	0.34

Table 5.14. Standard QC. Cases vs. WTCCC2 controls. The number of genes spanned by CNV events (GRate). Significance values of less than 0.05 are highlighted in bold.

Table 5.15 illustrates the number of CNV events to involve at least one gene. In this analysis PennCNV again supports the association of cases with rare deletion CNVs (p=0.0001), iPattern provides a non-statistically significant trend for support (p=0.07), also for duplications (p=0.057). QuantiSNP provides no such support.

	PennCNV			**iPattern**			**QuantiSNP**		
Event Frequency & Type	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.22 | 0.20	**0.0059**	**0.01**	0.16 | 0.15	0.14	0.27	0.15 | 0.15	0.84	0.35
Rare Del	0.16 | 0.12	**0.0001**	**0.0001**	0.096 | 0.086	0.07	0.14	0.068 | 0.075	0.88	0.29
Rare Dup	0.087 | 0.089	0.63	0.77	0.089 | 0.078	0.057	0.12	0.084 | 0.087	0.65	0.72

Table 5.15. Standard QC. Cases vs. WTCCC2 controls. The number of CNV events involving at least one gene (GProp). Significance values of less than 0.05 are highlighted in bold.

Finally, table 5.16 illustrates the number of genes per kilobase of CNV in cases and controls. In this instance the absolute figures are low. Nonetheless figures from the PennCNV method support the notion that rare deletions are more likely to involve genes in cases than controls (p=0.0099). iPattern and QuantiSNP do not support such an association (p=0.69 & 0.30 respectively).

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.0014 | 0.0013	0.17	0.34	0.0013 | 0.0013	0.41	0.83	0.0013 | 0.0013	0.75	0.50
Rare Del	0.0016 | 0.0014	**0.0099**	**0.02**	0.0013 | 0.0014	0.69	0.61	0.0014 | 0.0013	0.30	0.60
Rare Dup	0.0011 | 0.0011	0.68	0.66	0.0011 | 0.0011	0.54	0.93	0.0011 | 0.0011	0.64	0.69

Table 5.16. Standard QC. Cases vs. WTCCC2 controls. Number of genes per total CNV kb (GRich). Significance values of less than 0.05 are highlighted in bold.

5.4.1.2 High QC Threshold Results

SUMMARY: Results from both the PennCNV and iPatterns methods support our original hypothesis, however the degree of association through various metrics is often reduced, suggesting that some of our original association was being driven by samples at the cut-off threshold of our standard QC range. As in our standard QC analyses, results from the QuantiSNP method do not support conclusions drawn with the PennCNV and iPattern methods.

We next performed an identical analysis but this time in only those samples that passed our high QC thresholds described in 5.3.1.4.

5.4.1.2.1 Cases Vs. Screened Controls

2,397 cases and 332 screened controls passed our high QC thresholds and were included in the following analyses.

We initially compared the results of each method in our cases and screened controls.

Table 5.17 illustrates the CNV event rate per sample. Both PennCNV and iPattern call a significantly higher number of deletion CNVs in cases than controls (p=0.002 & p=0.029 respectively). There is no significant difference when the QuantiSNP method is used, and again this method calls less CNVs than either iPattern or PennCNV.

	PennCNV			**iPattern**			**QuantiSNP**		
Event Frequency & Type	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.92 | 0.70	**0.0013**	**0.0067**	0.82 | 0.78	0.072	0.16	0.72 | 0.73	0.63	0.80
Rare Del	0.59 | 0.39	**0.002**	**0.01**	0.47 | 0.36	**0.029**	0.072	0.32 | 0.34	0.74	0.59
Rare Dup	0.45 | 0.42	0.33	0.63	0.50 | 0.44	0.19	0.38	0.45 | 0.47	0.69	0.67

Table 5.17. High QC. Cases vs. screened controls. CNV event rate per person (Rate). Significance values of less than 0.05 are highlighted in bold.

Table 5.18 illustrates the proportion of cases and controls to have at least 1 CNV event. In this analysis, in contrast to our standard QC analysis, none of the methods support an association with rare deletions and cases.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.52 | 0.49	0.17	0.32	0.49 | 0.49	0.56	0.95	0.50 | 0.51	0.66	0.73
Rare Del	0.35 | 0.32	0.18	0.37	0.30 | 0.30	0.44	0.85	0.27 | 0.28	0.77	0.51
Rare Dup	0.33 | 0.33	0.57	0.95	0.31 | 0.34	0.82	0.41	0.34 | 0.34	0.55	0.95

Table 5.18. High QC. Cases vs. screened controls. Proportion of cases/controls to have at least one CNV event (Prop). Significance values of less than 0.05 are highlighted in bold.

Table 5.19 illustrates the total CNV event distance per subject. Similar to our results in the standard QC threshold, both PennCNV and iPattern detect a significant difference between cases and screened controls (p=0.0058 & p=0.014 respectively), principally driven by rare deletions (p=0.0061 & p=0.0043 respectively) and also by rare duplications called by the iPattern method (p=0.031). The QuantiSNP method does not support this finding, although the absolute difference is in the same direction. In comparison to the results in the standard QC group, the absolute total CNV event difference spanned per subject has fallen quite considerably for deletions and duplications, suggesting that a proportion of calls in our standard QC threshold analysis were clustering in samples of worse QC.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	440 | 347	**0.0058**	**0.034**	461 | 363	**0.014**	0.053	403 | 365	0.19	0.36
Rare Del	361 | 258	**0.0061**	0.059	408 | 269	**0.0043**	**0.049**	290 | 249	0.22	0.38
Rare Dup	393 | 353	0.17	0.33	451 | 359	**0.031**	0.083	397 | 390	0.45	0.85

Table 5.19. High QC. Cases vs. screened controls. Total CNV event distance spanned per subject in kb (KbTot). Significance values of less than 0.05 are highlighted in bold.

Table 5.20 illustrates the average CNV event size per subject. Similar to our standard QC analyses, none of the three methods detect a significant difference between cases and controls, suggesting that, if they occur, rare CNVs are no more likely to be of different size in cases or controls.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	258 | 247	0.32	0.58	269 | 258	0.34	0.63	269 | 254	0.27	0.51
Rare Del	219 | 213	0.44	0.82	249 | 225	0.24	0.46	230 | 211	0.30	0.54
Rare Dup	289 | 276	0.33	0.62	286 | 281	0.46	0.87	288 | 286	0.48	0.93

Table 5.20. High QC. Cases vs. screened controls. Average CNV event size in kb per subject (KbAvg). Significance values of less than 0.05 are highlighted in bold.

Table 5.21 illustrates the number of genes spanned by CNVs. Similar to our standard QC results, but with a less significant association, the PennCNV method indicates that case subjects are significantly more likely to have a deletion CNV spanning a gene than controls (p=0.024). The iPattern method shows a non-significant trend for association (p=0.082), similar to our standard QC analyses, and the QuantiSNP method does not show an association.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.20 | 0.16	0.068	0.12	0.17 | 0.13	0.072	0.15	0.15 | 0.16	0.71	0.72
Rare Del	0.13 | 0.084	**0.024**	0.054	0.096 | 0.069	0.082	0.16	0.072 | 0.066	0.42	0.75
Rare Dup	0.088 | 0.081	0.40	0.78	0.098 | 0.072	0.13	0.27	0.092 | 0.099	0.69	0.72

Table 5.21. High QC. Cases vs. screened controls. The number of genes spanned by CNV events (GRate). Significance values of less than 0.05 are highlighted in bold.

Table 5.23 illustrates the number of CNVs involving at least one gene. Similar to our standard QC analyses, but with a less convincing level of association, the PennCNV method supports the association of rare deletion CNVs with genes in cases over controls (p=0.043), but this is not supported by the QuantiSNP method (p=0.50), and the iPattern result fails to reach statistical significance (p=0.13).

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.17 | 0.15	0.14	0.27	0.15 | 0.12	0.12	0.25	0.14 | 0.15	0.81	0.44
Rare Del	0.11 | 0.078	**0.043**	0.087	0.089 | 0.069	0.13	0.25	0.068 | 0.066	0.50	0.91
Rare Dup	0.080 | 0.075	0.43	0.83	0.083 | 0.069	0.22	0.40	0.083 | 0.090	0.72	0.67

Table 5.22. High QC. Cases vs. screened controls. The number of CNV events involving at least one gene (GProp). Significance values of less than 0.05 are highlighted in bold.

Finally, table 5.23 illustrates the number of genes within each kilobase of CNV.

No method supports a significant association in this metric.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.0014 | 0.0013	0.43	0.85	0.0013 | 0.0011	0.13	0.26	0.0012 | 0.0013	0.64	0.72
Rare Del	0.0015 | 0.0012	0.13	0.25	0.0015 | 0.0011	0.12	0.24	0.0014 | 0.0011	0.13	0.28
Rare Dup	0.0011 | 0.0010	0.35	0.69	0.0011 | 0.00095	0.29	0.57	0.0011 | 0.0012	0.54	0.96

Table 5.23. High QC. Cases vs. screened controls. Number of genes per total CNV kb (GRich). Significance values of less than 0.05 are highlighted in bold.

5.4.1.2.2 Cases Vs. WTCCC2 Controls

We then compared the results of each method in our high QC cases and WTCCC2 controls.

2,397 cases and 4,424 WTCCC2 controls passed our high QC dataset and were used in the following analyses.

Table 5.24 illustrates the CNV event rate per person. Similar to our standard QC analysis both PennCNV and iPattern support the association of rare deletions with cases (p=0.0001 & p=0.0001 respectively). Again, the QuantiSNP method does not support this (p=0.98), however calls fewer events than either PennCNV or iPattern. In comparison to our standard QC analysis the absolute event rate has fallen quite noticeably, especially when the figures for PennCNV are inspected.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	1.03 | 0.89	**0.0003**	**0.0003**	0.92 | 0.84	**0.0078**	**0.016**	0.76 | 0.76	0.55	0.90
Rare Del	0.67 | 0.51	**0.0001**	**0.0001**	0.60 | 0.44	**0.0001**	**0.0001**	0.34 | 0.37	0.98	0.058
Rare Dup	0.48 | 0.50	0.80	0.41	0.50 | 0.48	0.18	0.34	0.47 | 0.48	0.65	0.70

Table 5.24. High QC. Cases vs. WTCCC2 controls. CNV event rate per person (Rate). Significance values of less than 0.05 are highlighted in bold.

Table 5.25 illustrates the proportion of subjects to have at least one CNV event. Similar to our standard QC analyses, although in contrast to those analyses done in high QC with screened controls, the PennCNV and iPattern methods support the association between rare deletions and cases (p=0.0066 & p=0.0056), whilst the QuantiSNP method does not (p=0.98), with a non-significant trend in the opposite direction (p=0.055). Again there is a modest reduction in significance levels and absolute figures, especially with the PennCNV method in comparison to the standard QC data.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.55 | 0.52	**0.01**	**0.018**	0.51 | 0.49	**0.029**	0.054	0.51 | 0.50	0.25	0.46
Rare Del	0.36 | 0.33	**0.0066**	**0.016**	0.34 | 0.30	**0.0056**	**0.0097**	0.28 | 0.30	0.98	0.055
Rare Dup	0.35 | 0.35	0.57	0.87	0.32 | 0.30	0.12	0.23	0.35 | 0.34	0.29	0.58

Table 5.25. High QC. Cases vs. WTCCC2 controls. Proportion of cases/controls to have at least one CNV event (Prop). Significance values of less than 0.05 are highlighted in bold.

Table 5.26 illustrates the total CNV event distance in kilobases per subject. In this analysis the iPattern and PennCNV methods support the association between deletion CNVs and cases (p=0.0095 & p=0.0001 respectively), whereas the QuantiSNP method does not (p=0.40). This association is increased from that observed in the standard QC equivalent analysis.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	455 | 458	0.42	0.96	499 | 447	**0.012**	**0.02**	406 | 497	0.82	0.44
Rare Del	386 | 333	**0.0095**	**0.014**	499 | 364	**0.0001**	**0.0001**	295 | 291	0.40	0.83
Rare Dup	400 | 449	0.91	0.29	447 | 442	0.43	0.85	397 | 530	0.88	0.42

Table 5.26. High QC. Cases vs. WTCCC2 controls. Total CNV event distance spanned per subject in kb (KbTot). Significance values of less than 0.05 are highlighted in bold.

Table 5.27 illustrates the average CNV event size in kilobases per subject. As in our previous analysis with screened controls, there is no statistically significant difference between cases and WTCCC2 controls via any of the three methods.

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	253 | 261	0.81	0.38	267 | 256	0.13	0.25	265 | 273	0.71	0.56
Rare Del	220 | 226	0.73	0.54	254 | 239	0.091	0.19	233 | 233	0.49	1.00
Rare Dup	289 | 298	0.79	0.40	282 | 280	0.43	0.9	284 | 296	0.75	0.50

Table 5.27. High QC. Cases vs. WTCCC2 controls. Average CNV event size in kb per subject (KbAvg). Significance values of less than 0.05 are highlighted in bold.

Table 5.28 illustrates the number of genes spanned by CNV events in cases and WTCCC2 controls. PennCNV supports the association between number of genes

spanned by deletion CNVs and cases (p=0.0005). iPattern has a non-significant trend in the same direction (p=0.062). This is not supported by the QuantiSNP method (p=0.74).

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.22 | 0.20	**0.04**	0.074	0.18 | 0.17	0.15	0.27	0.16 | 0.17	0.89	0.24
Rare Del	0.14 | 0.10	**0.0005**	**0.0008**	0.11 | 0.094	0.062	0.11	0.072 | 0.076	0.74	0.55
Rare Dup	0.088 | 0.099	0.91	0.19	0.098 | 0.087	0.12	0.23	0.092 | 0.10	0.85	0.33

Table 5.28. High QC. Cases vs. WTCCC2 controls. The number of genes spanned by CNV events (GRate). Significance values of less than 0.05 are highlighted in bold.

Table 5.29 illustrates the number of CNV events to involve at least one gene. In this analysis PennCNV and iPattern support the association of cases with rare deletion CNVs involving genes (p=0.0002 & p=0.036 respectively). QuantiSNP does not provide support for this conclusion however (p=0.76).

	PennCNV			iPattern			QuantiSNP		
Event Frequency & Type	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.19 | 0.17	**0.027**	**0.047**	0.16 | 0.15	0.15	0.27	0.14 | 0.15	0.83	0.38
Rare Del	0.12 | 0.090	**0.0002**	**0.0002**	0.097 | 0.084	**0.036**	0.068	0.068 | 0.073	0.76	0.49
Rare Dup	0.080 | 0.088	0.88	0.27	0.083 | 0.077	0.17	0.31	0.083 | 0.088	0.75	0.53

Table 5.29. High QC. Cases vs. WTCCC2 controls. The number of CNV events involving at least one gene (GProp). Significance values of less than 0.05 are highlighted in bold.

Finally, table 5.30 illustrates the number of genes per kilobase of CNV in cases and controls. In this instance the absolute figures are again low. The PennCNV method supports the notion that rare deletions are more likely to involve genes in cases than controls (p=0.044). iPattern and QuantiSNP do not support such an association (p=0.58 & p=0.15 respectively).

	PennCNV			**iPattern**			**QuantiSNP**		
Event Frequency & Type	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare Del/Dup	0.00137 | 0.00136	0.42	0.86	0.00128 | 0.00125	0.35	0.70	0.00125 | 0.00129	0.67	0.65
Rare Del	0.00149 | 0.00130	**0.044**	0.088	0.00132 | 0.00135	0.58	0.82	0.00138 | 0.00126	0.15	0.30
Rare Dup	0.00104 | 0.00116	0.87	0.26	0.00107 | 0.00109	0.55	0.86	0.00107 | 0.00115	0.77	0.45

Table 5.30. High QC. Cases vs. WTCCC2 controls. Number of genes per total CNV kb (GRich). Significance values of less than 0.05 are highlighted in bold.

5.4.2 Results from the Intersected Call Set

SUMMARY: Within the intersected call set we failed to see any evidence of association between the case cohort with any CNV burden metric. This is expected as all calls from this dataset must be derived from all three methods, and results from the QuantiSNP method alone did not yield significant results.

Subsequently we performed a similar analysis on the intersected call set described in 5.3.1.3. Since the QuantiSNP method did not show positive results in our single method analyses, it would be reasonable to expect that an intersected call set might also show this. For brevity we have presented these analyses in the appendix.

5.4.3 Validation of Calls Made Over 22q11.2 Made with Each Method

Our analyses above lead to a confusing picture. Are we observing a true difference in CNV burden between cases and controls, or are our results confounded by false positive calls made by some methods but not others? In an attempt to answer this question we took our array CGH data from chapter 4, and validated all calls of >100kb in length and made with ≥10 SNPs that fell within the 22q11.2 region covered by our CGH array, from each method separately. We then calculated the number of validated calls as a percentage of the total number of calls made by each method.

Table 5.31 illustrates the results of our validation.

95.8% of PennCNV calls and 95.8% QuantiSNP calls are confirmed, with 81.3% of iPattern calls confirmed. Of those calls that are not confirmed, there is only 1 deletion called by PennCNV, 9 by iPattern and 2 by QuantiSNP.

Method	Replication Status	N Dels | Dups | Total	% Dels | Dups | Total
PennCNV	Replicated	39 | 29 | 68	54.9 | 40.8 | 95.8
	Not Replicated	1 | 2 | 3	1.4 | 2.8 | 4.2
	Total	40 | 31 | 71	56.3 | 43.7 | 100.0
iPattern	Replicated	36 | 25 | 61	48.0 | 33.3 | 81.3
	Not Replicated	9 | 5 | 14	12.0 | 6.7 | 18.7
	Total	45 | 30 | 75	60.0 | 40.0 | 100.0
QuantiSNP	Replicated	38 | 30 | 68	53.5 | 42.3 | 95.8
	Not Replicated	2 | 1 | 3	2.8 | 1.4 | 4.2
	Total	40 | 31 | 71	56.3 | 43.7 | 100.0

Table 5.31. Validation of calls (≥10 markers, >100kb) from the three methods used in our analyses over 22q11.2.

This data fails to clarify our original question. Calls within the 22q11.2 region are well validated with both PennCNV and QuantiSNP, but these methods do not give results that support each other in our analyses presented in 5.4.1. The implication from this is that the calls differing between the two methods do not fall within this region of the genome

5.4.4 Follow Up of Regions Previously Associated with Schizophrenia in our High QC Intersected Call Set

We originally reported frequencies of CNVs in regions previously implicated in schizophrenia in our PennCNV calls (see 2.4.5). We now revisit these regions with our high QC, intersected data. To improve power to detect differences we included 424 screened control samples derived from cheek swab DNA that we had previously excluded (see 2.3.4.1) but which did pass our high QC thresholds.

These samples were not used in genome wide burden calculations, but can reasonably be included here since the CNVs in these regions are large and easily validated by visual inspection of plots. Associated screenshots from the UCSC browser are reproduced in Figs. 5.9, 5.10, 5.11, 5.12, 5.13, 5.14 & 5.15. We tested the number of CNVs in each region in cases versus each control cohort with Fisher's exact tests, shown in table 5.32. There were no significant differences between cohorts, however we noted with some surprise that large CNVs, especially duplications, had been removed from this dataset. On further investigation this was not necessarily due to the samples being removed through our high QC thresholds, but rather these CNVs had never been called by the iPattern and QuantiSNP methods (Fig. 5.16)

Locus & Position (MB)	CNV Region Cov'ge	Number of Samples with CNVs			Cases vs. WT2 P val, OR (95%CI)	Cases vs. Scr Cont P val, OR (95%CI)
		Cases N=2397	WT2 Controls N=4424	Scr'ned Controls N=756		
1q21.1 144.9-146.3	Del ≥90%	2	3	0	0.57, 1.23 (0-6.16)	0.58, NaN (0.16-NaN)
	Del <90%	2	1	2	0.28, 3.69 (0.48-NaN)	0.24, 0.31 (0.06-1.79)
	Total	4	4	2	0.30, 1.85 (0.51-6.74)	0.44, 0.63 (0.13-NaN)
	Dup ≥90%	0	2	0	0.42, 0 (0-3.55)	N/A
	Dup <90%	2	2	0	0.44, 1.85 (0.33-10.48)	0.58, NaN (0.16-NaN)
	Total	2	4	0	0.64, 0.92 (0-4.31)	0.58, NaN (0.16-NaN)
2p16.3 50.0-51.2	Del ≥90%	0	0	0	N/A	N/A
	Del <90%	5	16	3	0.20, 0.58 (0.22-1.52)	0.30, 0.52 (0.14-1.99)
	Total	5	16	3	0.20, 0.58 (0.22-1.52)	0.30, 0.52 (0.14-1.99)
	Dup ≥90%	0	0	0	N/A	N/A
	Dup <90%	0	0	0	N/A	N/A
	Total	0	0	0	N/A	N/A
15q11.2 20.2-20.8	Del ≥90%	10	18	4	0.55, 1.03 (0.48-2.19)	0.44, 0.79 (0.26-2.38)
	Del <90%	0	1	0	0.65, 0 (0-NaN)	N/A
	Total	10	19	4	0.56, 0.97 (0.46-2.06)	0.44, 0.79 (0.26-2.38)
	Dup ≥90%	9	13	6	0.36, 1.28 (0.56-2.93)	0.13, 0.47 (0.17-1.27)
	Dup <90%	5	2	0	0.06, 4.62 (1.03-NaN)	0.25, NaN (0.41-NaN)
	Total	14	15	6	0.10, 1.73 (0.84-3.53)	0.34, 0.73 (0.29-1.86)
15q13.3 29.0-30.3	Del ≥90%	0	0	0	N/A	N/A
	Del <90%	0	0	0	N/A	N/A
	Total	0	0	0	N/A	N/A
	Dup ≥90%	12	35	6	0.11, 0.63 (0.33-1.2)	0.25, 0.63 (0.24-1.62)
	Dup <90%	1	2	0	0.72, 0.92 (0-7.04)	0.76, NaN (0-NaN)
	Total	13	37	6	0.11, 0.65 (0.35-1.21)	0.29, 0.68 (0.27-1.74)
16p13.1 15.0-16.4	Del ≥90%	3	3	0	0.36, 1.85 (0.43-8.00)	0.44, NaN (0.25-NaN)
	Del <90%	1	2	0	0.72, 0.92 (0-7.04)	0.76, NaN (0-NaN)
	Total	4	5	0	0.39, 1.48 (0.43-5.08)	0.33, NaN (0.33-NaN)
	Dup ≥90%	2	7	0	0.33, 0.53 (0-2.23)	0.58, NaN (0.16-NaN)
	Dup <90%	2	2	0	0.44, 1.85 (0.33-10.48)	0.58, NaN (0.16-NaN)
	Total	4	9	0	0.50, 0.82 (0.27-2.51)	0.33, NaN (0.33-NaN)
16p11.2 29.5-30.1	Del ≥90%	1	2	0	0.72, 0.92 (0-7.04)	0.76, NaN (0-NaN)
	Del <90%	0	0	0	N/A	N/A
	Total	1	2	0	0.72, 0.92 (0-7.04)	0.76, NaN (0-NaN)
	Dup ≥90%	0	0	0	N/A	N/A
	Dup <90%	0	0	0	N/A	N/A
	Total	0	0	0	N/A	N/A
22q11.2 17.4-19.8	Del ≥90%	1	0	0	0.35, NaN (0-NaN)	0.76, NaN (0-NaN)
	Del <90%	1	0	0	0.35, NaN (0-NaN)	0.76, NaN (0-NaN)
	Total	2	0	0	0.12, NaN (0.96-NaN)	0.58, NaN (0.16-NaN)
	Dup ≥90%	1	2	0	0.72, 0.92 (0-7.04)	0.76, NaN (0-NaN)
	Dup <90%	3	8	0	0.42, 0.69 (0.20-2.40)	0.44, NaN (0.25-NaN)
	Total	4	10	0	0.42, 0.74 (0.24-2.23)	0.33, NaN (0.33-NaN)
All Regions	Del total	17	26	4	0.32, 1.21 (0.66-2.21)	0.41, 1.34 (0.47-3.82)
	Dup total	24	59	12	0.14, 0.75 (0.47-1.20)	0.13, 0.63 (0.32-1.24)
	All total	41	85	16	0.30, 0.89 (0.61-1.29)	0.28, 0.8 (0.45-1.43)

Table 5.32. Numbers of samples with CNVs in our high QC intersected call set, stratified by type, in genomic regions previously implicated in schizophrenia. P values are calculated with Fisher's exact method. NaN - not a number.

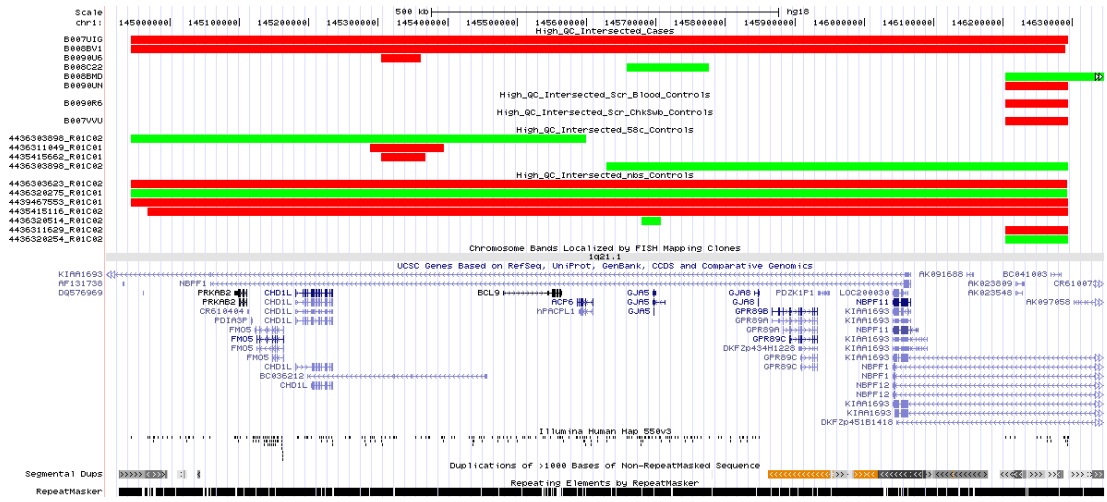


Fig. 5.9. 1q21.1 region. Red lines indicate deletions, green lines duplications. Cases appear above screened controls, which appear above WTCCC2 controls.

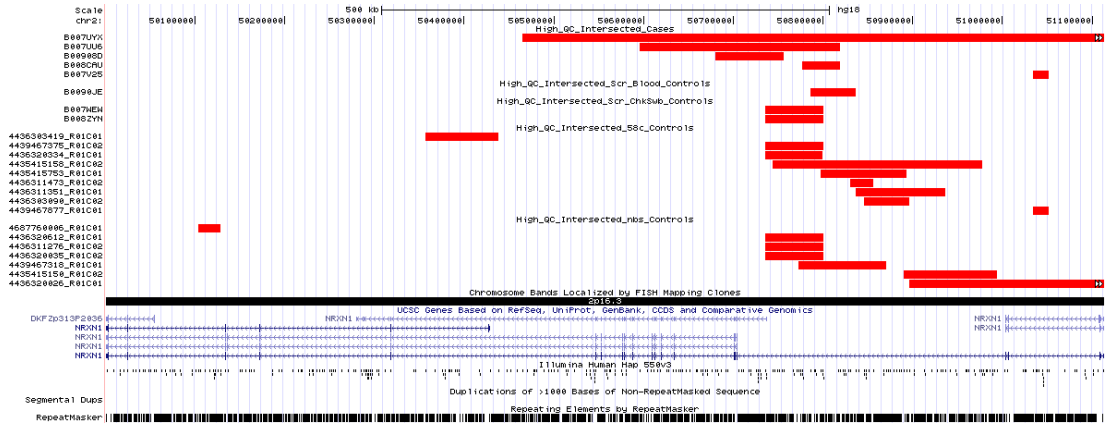


Fig. 5.10. 2p16.3 region. Red lines indicate deletions, green lines duplications. Cases appear above screened controls, which appear above WTCCC2 controls.

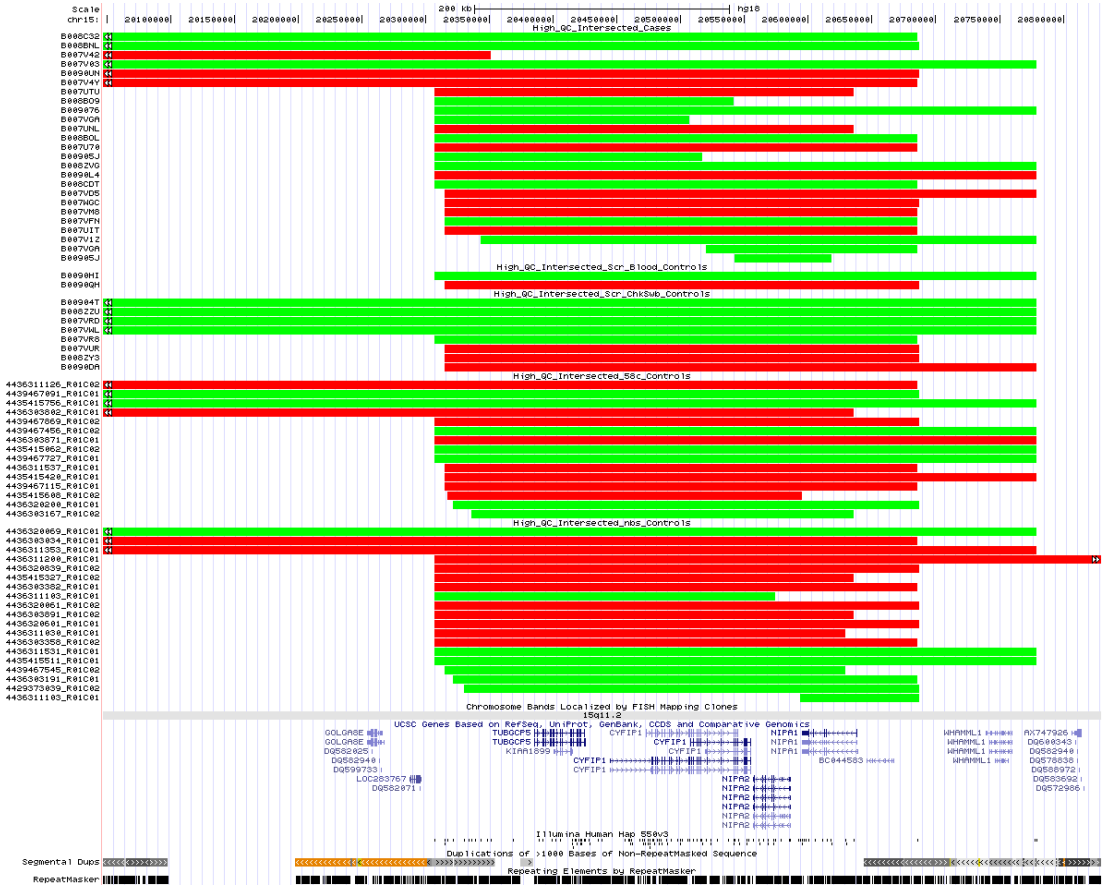


Fig. 5.11. 15q11.2 region. Red lines indicate deletions, green lines duplications. Cases appear above screened controls, which appear above WTCCC2 controls.

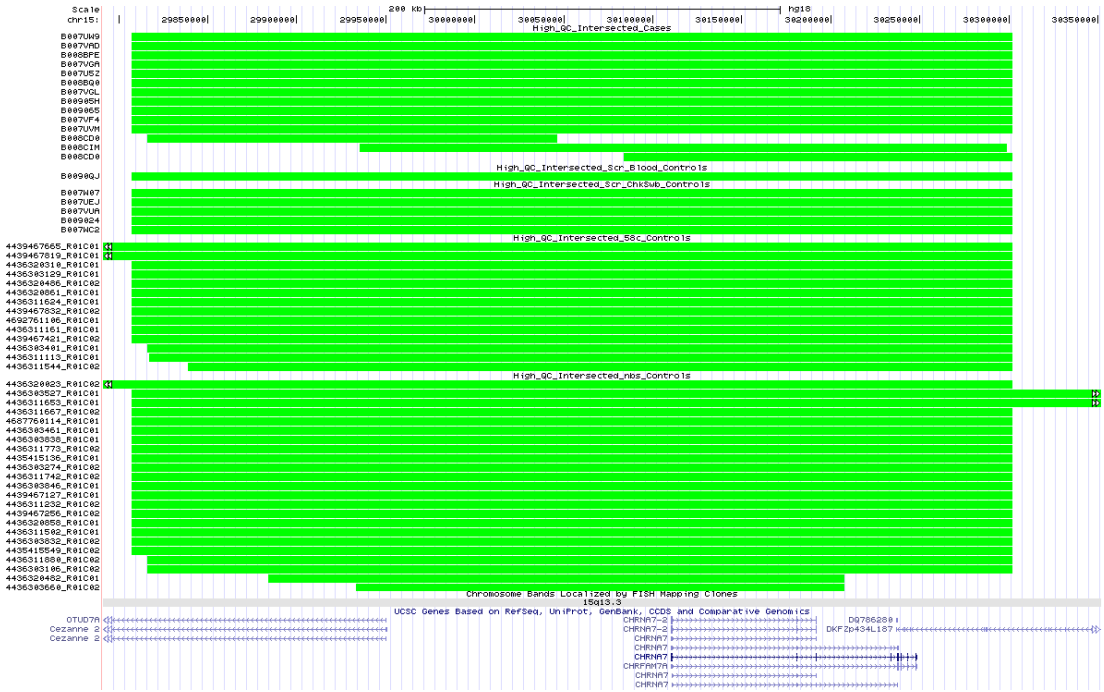


Fig. 5.12. 15q13.3 region. Red lines indicate deletions, green lines duplications. Cases appear above screened controls, which appear above WTCCC2 controls.

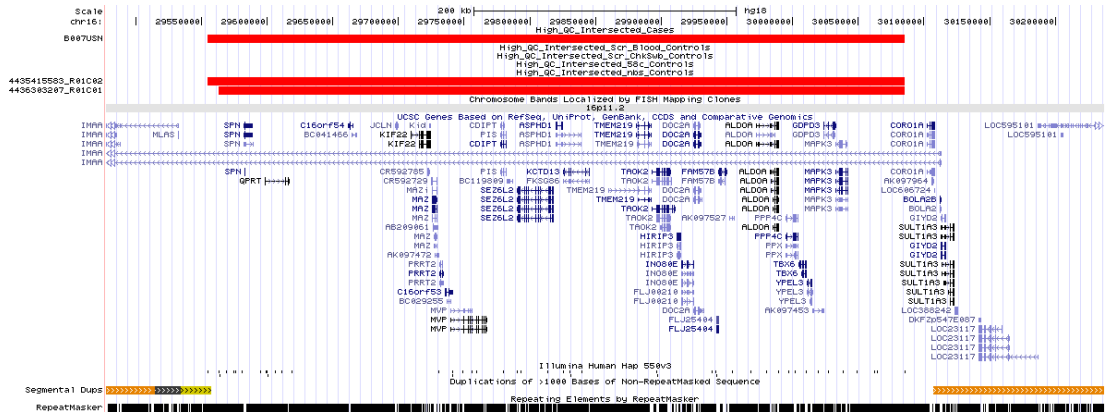


Fig. 5.13. 16p11.2 region. Red lines indicate deletions, green lines duplications. Cases appear above screened controls (no CNVs seen), which appear above WTCCC2 controls.

Fig. 5.14. 16p13.1 region. Red lines indicate deletions, green lines duplications. Cases appear above screened controls (no CNVs seen), which appear above WTCCC2 controls.

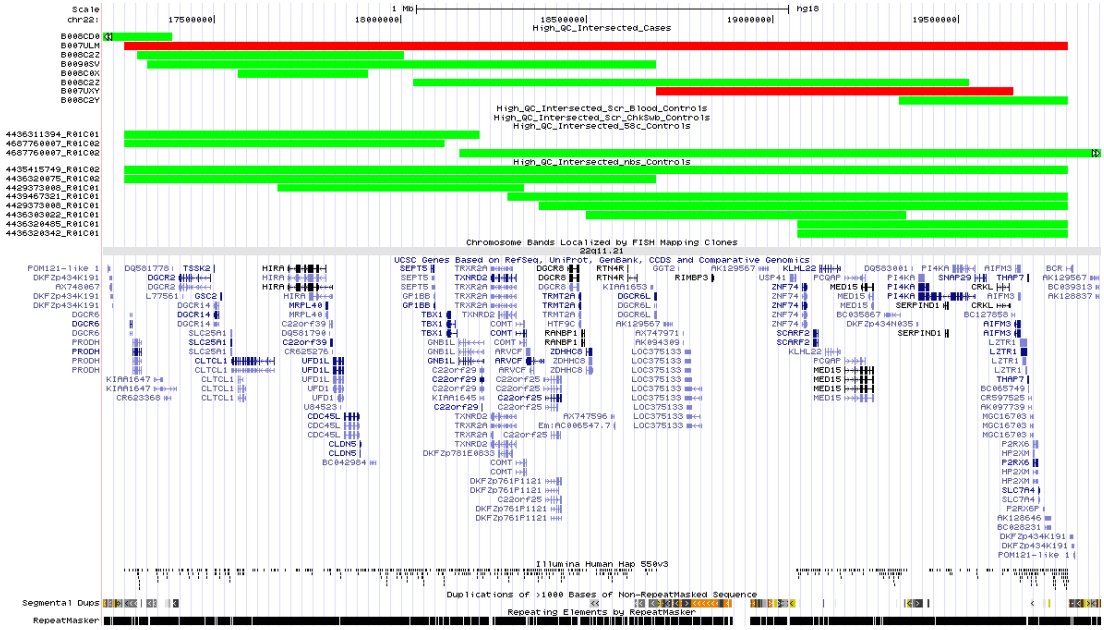


Fig. 5.15. 22q11.2 region. In this plot relatively common CNVs within the gene PRODH have been removed for conciseness. Red lines indicate deletions, green lines duplications. Cases appear above screened controls (no CNVs seen), which appear above WTCCC2 controls.

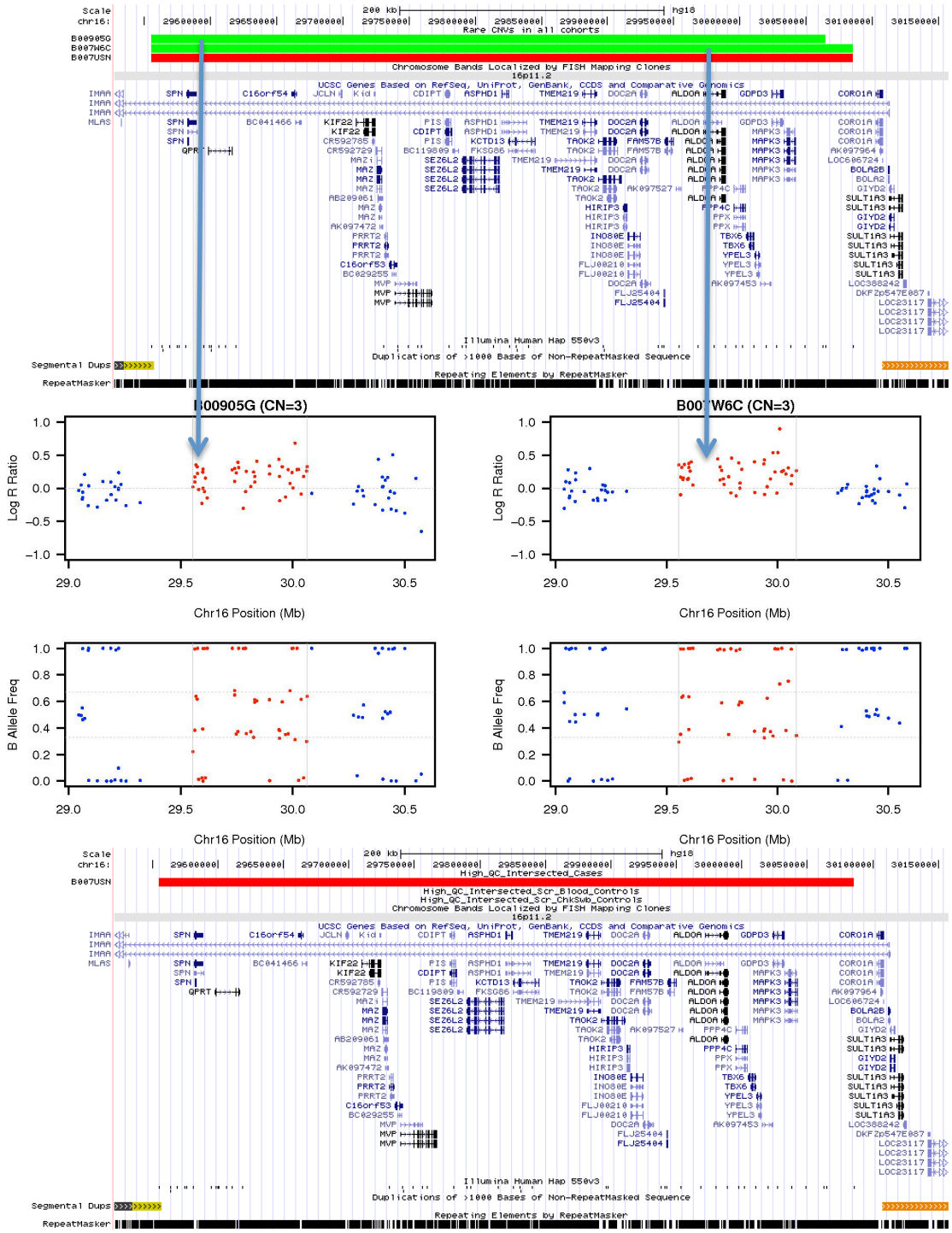
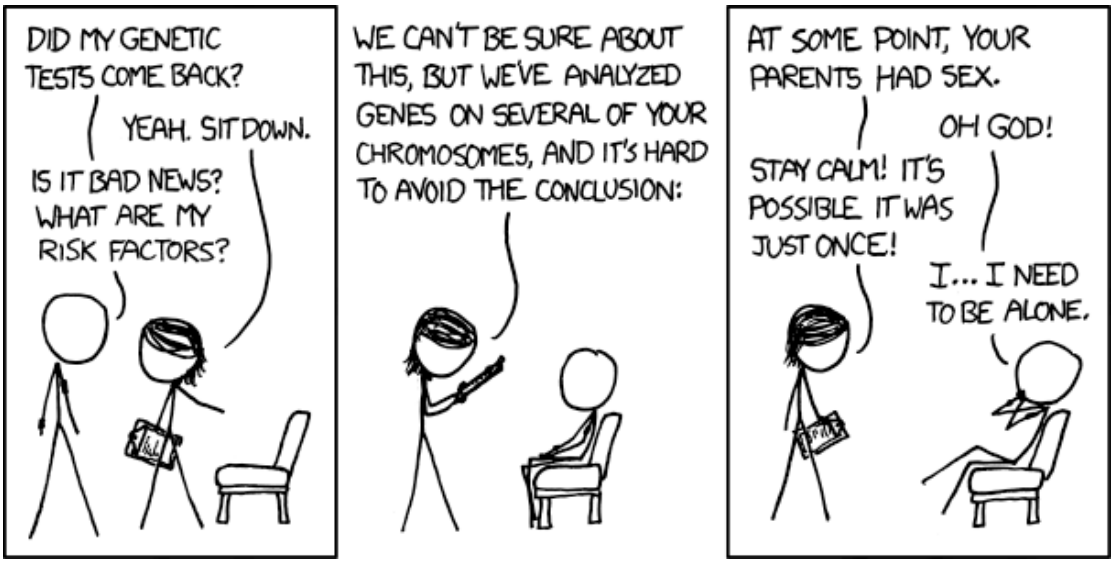


Fig. 5.16. Two large duplications in 16p11.2, called by PennCNV in our high QC dataset (green lines, upper panel), and clearly visible on visual QC of the plots (middle panel), fail to be called by the QuantiSNP and iPattern methods and are consequently absent from our high QC intersected dataset (lower panel)

5.5 Conclusion

We originally hypothesised that the association between cases and rare deletion CNVs would be retained by using data derived from QuantiSNP, iPattern and a set of consensus calls made from intersecting calls from all three methods. Whilst reanalysis of our data with the iPattern method provided additional support for our association described in chapter 2, reanalysis with the QuantiSNP method failed to provide additional support. A reanalysis of our data with a sample set derived of only high QC samples reduced the level of association in most metrics. An analysis using an intersected call set made from calls derived from all three methods did not support our original associations, although this call set may be overly conservative. Follow up of calls made by all three methods in the 22q11.2 region suggested that the PennCNV and QuantiSNP methods performed better than the iPattern method, although all methods had a high rate of validation, and the rate of validation over the 22q11.2 region may not reflect the validation rate overall. A reanalysis of regions previously implicated in schizophrenia revealed no significant differences between case and control populations, although some large CNVs appear not to have been called by some calling methods, leading to an over-conservative call set.

Chapter 6. Phenotype Analyses, Sex Chromosome Syndromes and Specific CNVs



6.1 Introduction

In this final analysis chapter we take our CNV call sets from different methods and run association analyses with various phenotypic datasets available from our cohorts, namely age of onset of disorder, duration of worst episode, personality trait scores (psychoticism, neuroticism and extraversion) and heritable factor scores derived from mood-related SCAN items analysed previously in work by Korszun et al(Korszun et al., 2004). We were particularly interested to study CNV deletion burden in relation to age of onset of disorder, as some studies have highlighted an increased burden of disorder in cohorts with earlier age of onset of psychiatric disorder(Malhotra et al., 2011). Finally we describe an analysis of the frequency of sex chromosome abnormalities across our data and then present a case series of interesting CNVs in our case cohort, along with psychometric data where they were available.

6.2 Hypotheses

A) Higher CNV deletion burden will be related to lower age of onset of disorder

B) There will be a relationship between CNV burden and symptom pattern and/or personality traits.

6.3 Methods

In this analysis we analysed a total of 1,940 samples from our case cohort only. All samples were derived from our high QC cohort described in 5.3.1.4. We used

CNV calls derived from PennCNV, iPattern, QuantiSNP and the intersected call set derived from all three methodologies.

We performed association analyses across cohorts using phenotypic data that had been collected across cohorts, specifically

a) Age of first onset of disorder
b) Duration of worst episode
c) Trait neuroticism scores
d) Trait psychoticism scores
e) Trait extraversion scores
f) Factors from research into the familiality of symptom dimensions in depression(Korszun et al., 2004), namely
    i) Factor 1 - Mood symptoms. Derived from SCAN items such as 6.001 Depressed mood, 6.004 Anhedonia and 6.006 Loss of hope for the future.
    ii) Factor 2 - Guilt and psychomotor agitation. Derived from SCAN items such as 6.010 Preoccupation with death or catastrophe, 6.013 Pathological guilt and 6.014 Guilty ideas of reference.
    iii) Factor 3 - Atypical depressive features. Derived from SCAN items such as 6.009 Morning depression, 8.007 Weight gain and 8.016 Hypersomnia.

The phenotypic data from across the DeNT, DeCC and GENDEP studies has been previously collated and curated into a single database (Butler, Cohen-Woods, Farmer, McGuffin, & Lewis, 2010). We extracted data from this, querying the

items above. Trait personality scores are derived from the Eysenck Personality Questionnaire(H. Eysenck & Eysenck, 1964), the results of which place individuals on continua comprised of psychoticism, neuroticism and extraversion.

Factors from Korszun et al's research into the familiality of depression symptoms are derived from SCAN items(Korszun et al., 2004). Since not all SCAN items were used in the studies contributing to the GWAS, we were able to re-construct three out of four factors, as detailed above, but we were unable to reconstruct the anxiety factor.

We restricted our sample set in each analysis to those samples where phenotype data was not missing. We analysed our phenotypes with our call sets derived from each of the three separate methods and the call set derived from the intersection of all three methods. We analysed two CNV burden metrics:

1. The number of rare CNVs per person

2. The total burden of rare CNVs in kb per person

We then stratified these two datasets into deletion CNVs, duplication CNVs and all CNVs together and compared each of these datasets with each phenotype using linear bootstrapped regression analysis with 1,000 replications corrected for the LRRSD, centre of recruitment, age and sex using STATA IC v10.1. Bootstrapping allows association analyses between variables when the data do not conform to a standard distribution(Wu, 1986). PLINK currently does not

have the functionality to calculate significance levels for quantitative traits in a genome-wide manner with CNV data, hence our use of this method.

We undertook 24 separate tests for each phenotype, which included 3 CNV types (deletions, duplications and all CNVs), 4 method sets (PennCNV, iPattern, QuantiSNP and the intersected call set from all three methods) and 2 ways of representing CNV burden (number of CNVs per person and total burden in kb). Therefore we set a Bonferroni correction for multiple testing as 0.05/24=0.0021.

6.4 Results

6.4.1 Phenotypic Association Analyses

SUMMARY: Whilst some trends were observed ($p<0.05$), no test for association survived Bonferroni correction for multiple testing ($p<0.0021$). The most interesting trend observed was between deletion burden and trait neuroticism.

We report, in each instance, the number of phenotypic observations, the z score and the probability that the z score occurred due to chance.

6.4.1.1 Age of Onset

Fig. 6.1 illustrates the distribution of age of onset.

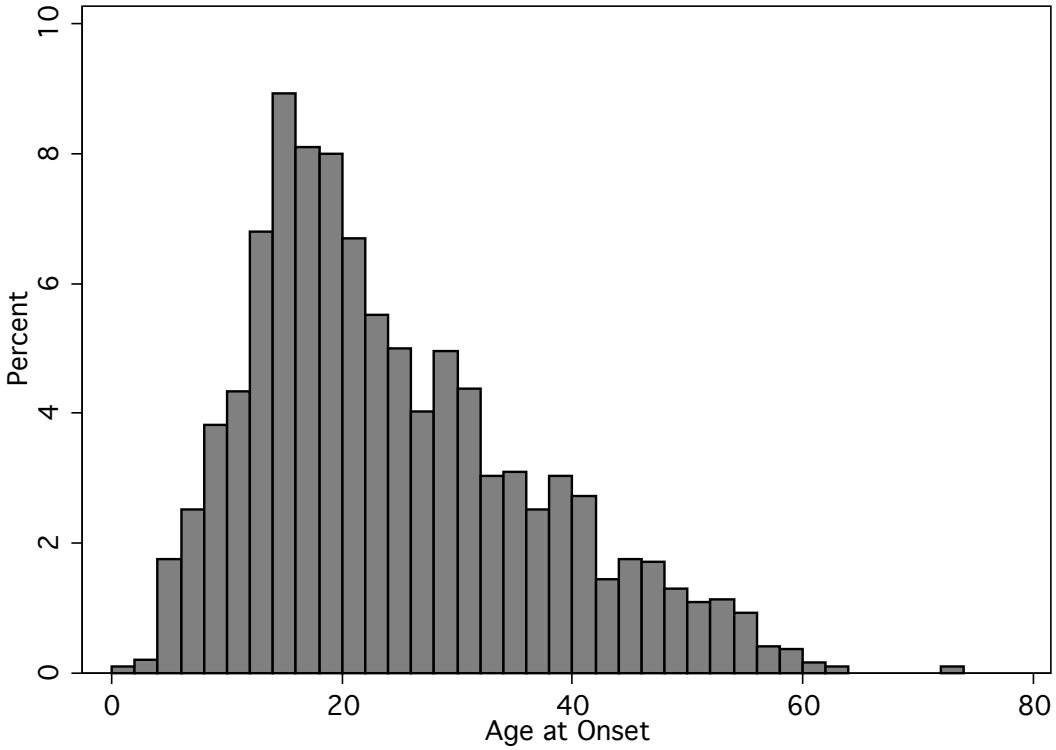


Fig. 6.1. Histogram of age of onset of depressive disorder.

Table 6.1 illustrates the results of bootstrapped linear regression analysis for the number of CNVs and total length of CNVs per person compared to age of onset of depressive disorder (log transformed) corrected for age, centre of recruitment, sex, and LRRSD. There are no significant associations within this group.

Phenotype	CNV Call Set	CNV Type	CNV Metric	No. of Obs.	z	P>|z|
Age of Onset	Intersected	All	No. of CNVs	1926	2.64	0.008
			CNV Burden (Kb)	1926	1.62	0.105
		Deletions	No. of CNVs	1926	1.63	0.103
			CNV Burden (Kb)	1926	1.18	0.238
		Duplications	No. of CNVs	1926	2.03	0.042
			CNV Burden (Kb)	1926	1.20	0.231
	PennCNV	All	No. of CNVs	1926	2.11	0.035
			CNV Burden (Kb)	1926	1.87	0.062
		Deletions	No. of CNVs	1926	1.78	0.076
			CNV Burden (Kb)	1926	1.68	0.093
		Duplications	No. of CNVs	1926	1.22	0.222
			CNV Burden (Kb)	1926	1.09	0.274
	iPattern	All	No. of CNVs	1926	0.62	0.536
			CNV Burden (Kb)	1926	0.88	0.380
		Deletions	No. of CNVs	1926	1.02	0.308
			CNV Burden (Kb)	1926	1.14	0.254
		Duplications	No. of CNVs	1926	0.13	0.894
			CNV Burden (Kb)	1926	0.36	0.719
	QuantiSNP	All	No. of CNVs	1926	0.52	0.602
			CNV Burden (Kb)	1926	0.74	0.457
		Deletions	No. of CNVs	1926	-0.82	0.414
			CNV Burden (Kb)	1926	-0.34	0.737
		Duplications	No. of CNVs	1926	1.17	0.241
			CNV Burden (Kb)	1926	1.09	0.278

Table 6.1. Association analysis of CNV burden called with different methods with age of onset (log transformed). Bonferroni adjusted p value for significance = 0.0021.

6.4.1.2 Duration of Worst Episode

Fig. 6.2 illustrates the distribution of duration of worst depressive episode. No specific transformation improved the distribution of this metric.

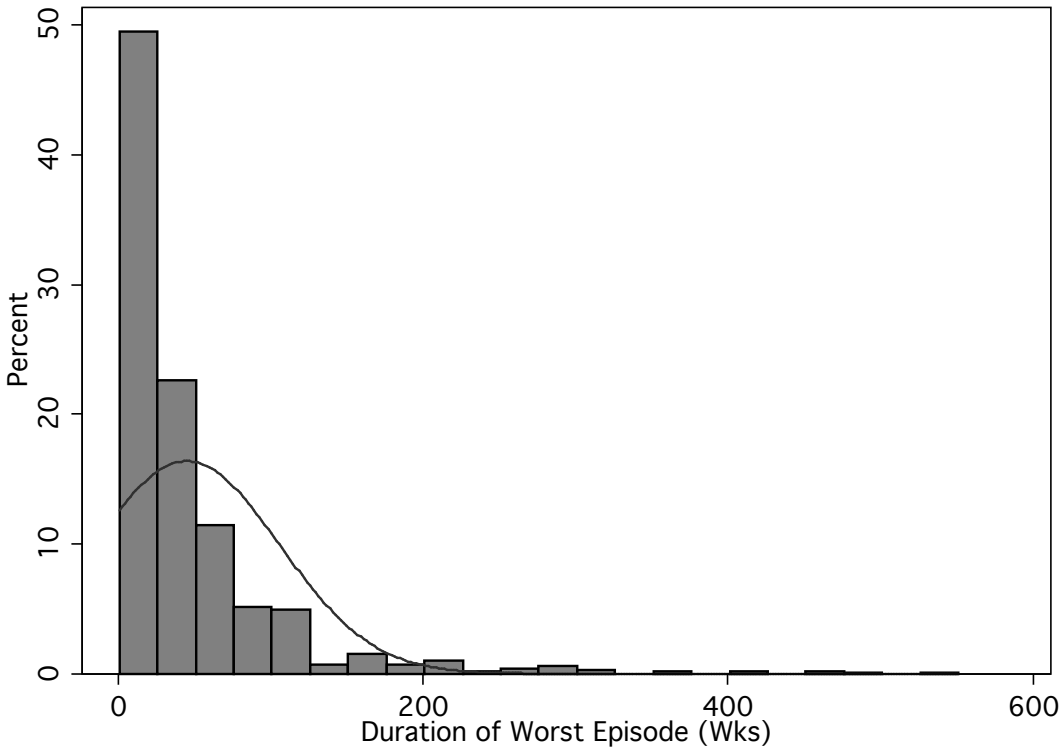


Fig. 6.2. Histogram of duration of worst depressive episode, with a superimposed normal distribution plot.

Table 6.2 illustrates the results of bootstrapped linear regression analysis for the number of CNVs and total length of CNVs per person compared to duration of worst depressive episode corrected for age, centre of recruitment, sex, and LRRSD. There are no significant associations within this group.

Phenotype	CNV Call Set	CNV Type	CNV Metric	No. of Obs.	z	P>|z|
Duration of Worst Episode	Intersected	All	No. of CNVs	977	-0.30	0.764
			CNV Burden (Kb)	977	-0.30	0.765
		Deletions	No. of CNVs	977	-0.30	0.767
			CNV Burden (Kb)	977	-0.30	0.767
		Duplications	No. of CNVs	977	-0.30	0.761
			CNV Burden (Kb)	977	-0.29	0.772
	PennCNV	All	No. of CNVs	977	-0.30	0.765
			CNV Burden (Kb)	977	-0.30	0.761
		Deletions	No. of CNVs	977	-0.30	0.762
			CNV Burden (Kb)	977	-0.30	0.765
		Duplications	No. of CNVs	977	-0.30	0.763
			CNV Burden (Kb)	977	-0.30	0.766
	iPattern	All	No. of CNVs	977	-0.30	0.765
			CNV Burden (Kb)	977	-0.30	0.764
		Deletions	No. of CNVs	977	-0.30	0.762
			CNV Burden (Kb)	977	-0.29	0.772
		Duplications	No. of CNVs	977	-0.28	0.778
			CNV Burden (Kb)	977	-0.30	0.763
	QuantiSNP	All	No. of CNVs	977	-0.30	0.761
			CNV Burden (Kb)	977	-0.30	0.765
		Deletions	No. of CNVs	977	-0.29	0.769
			CNV Burden (Kb)	977	-0.30	0.764
		Duplications	No. of CNVs	977	-0.29	0.770
			CNV Burden (Kb)	977	-0.30	0.765

Table 6.2. Association analysis of CNV burden called with different methods with the duration of worst depressive episode. Bonferroni adjusted p value for significance = 0.0021.

6.4.1.3 Factor Analyses

6.4.1.3.1 Factor 1 - Mood Symptoms

Fig. 6.3 illustrates the distribution of mood symptoms, as defined by dimensions of SCAN symptoms in Korszun et al(Korszun et al., 2004).

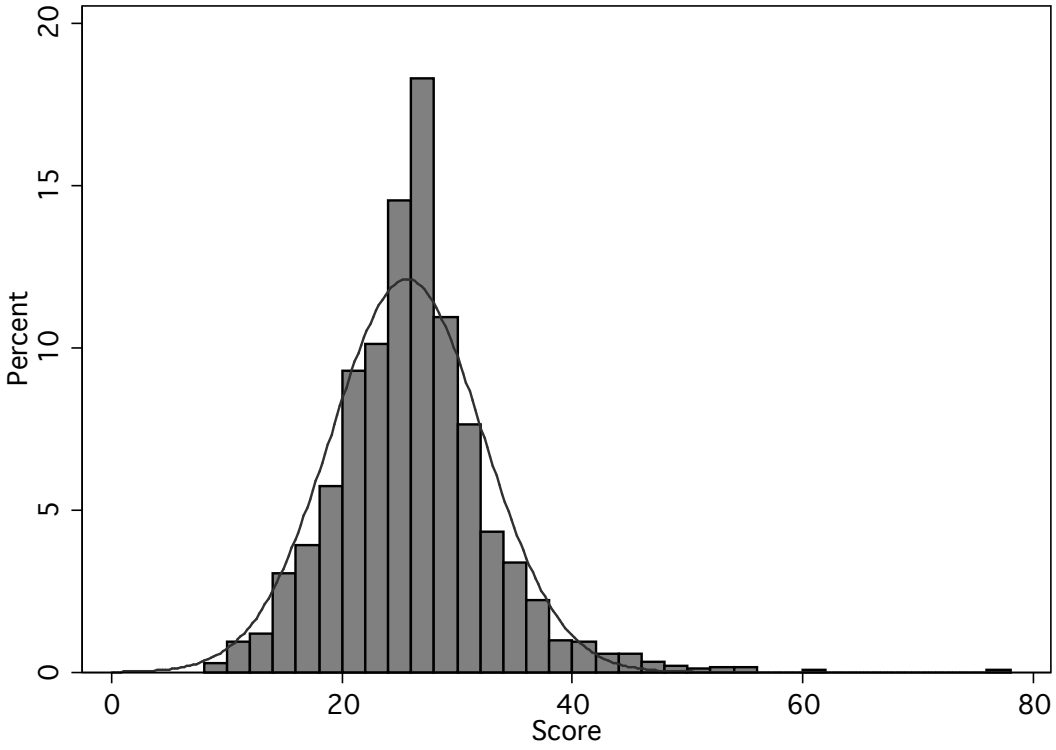


Fig. 6.3. Histogram of the mood symptoms dimension with a superimposed normal distribution plot.

Table 6.3 illustrates the results of bootstrapped linear regression analysis for the number of CNVs and total length of CNVs per person compared to the mood symptoms dimension corrected for age, centre of recruitment, sex, and LRRSD. There are no significant associations within this group.

Phenotype	CNV Call Set	CNV Type	CNV Metric	No. of Obs.	z	P>|z|
Factor 1	Intersected	All	No. of CNVs	1924	0.82	0.410
			CNV Burden (Kb)	1924	2.20	0.028
		Deletions	No. of CNVs	1924	-0.55	0.580
			CNV Burden (Kb)	1924	1.43	0.154
		Duplications	No. of CNVs	1924	1.48	0.139
			CNV Burden (Kb)	1924	1.75	0.079
	PennCNV	All	No. of CNVs	1924	-0.61	0.540
			CNV Burden (Kb)	1924	1.10	0.269
		Deletions	No. of CNVs	1924	-1.80	0.071
			CNV Burden (Kb)	1924	0.33	0.744
		Duplications	No. of CNVs	1924	1.18	0.237
			CNV Burden (Kb)	1924	1.29	0.198
	iPattern	All	No. of CNVs	1924	0.31	0.755
			CNV Burden (Kb)	1924	1.55	0.122
		Deletions	No. of CNVs	1924	-2.32	0.020
			CNV Burden (Kb)	1924	0.17	0.867
		Duplications	No. of CNVs	1924	0.72	0.471
			CNV Burden (Kb)	1924	0.83	0.404
	QuantiSNP	All	No. of CNVs	1924	1.30	0.195
			CNV Burden (Kb)	1924	2.09	0.037
		Deletions	No. of CNVs	1924	0.34	0.731
			CNV Burden (Kb)	1924	1.57	0.117
		Duplications	No. of CNVs	1924	1.34	0.179
			CNV Burden (Kb)	1924	1.55	0.120

Table 6.3. Association analysis of CNV burden called with different methods with the mood symptoms dimension. Bonferroni adjusted p value for significance = 0.0021.

6.4.1.3.2 Factor 2 - Guilt and Psychomotor Agitation

Fig. 6.4 illustrates the distribution of guilt and psychomotor agitation symptoms, as defined by dimensions of SCAN symptoms in Korszun et al(Korszun et al., 2004).

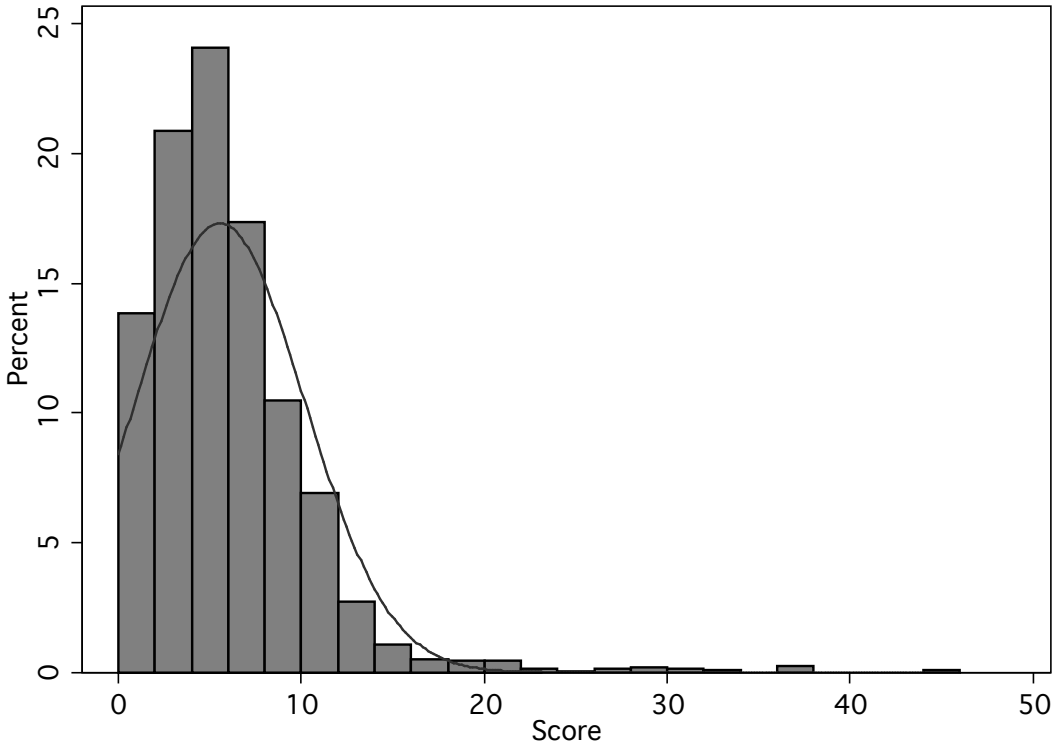


Fig. 6.4. Histogram of the guilt and psychomotor agitation symptoms dimension with a superimposed normal distribution plot.

Table 6.4 illustrates the results of bootstrapped linear regression analysis for the number of CNVs and total length of CNVs per person compared to the guilt and psychomotor agitation symptoms dimension corrected for age, centre of recruitment, sex, and LRRSD. There are no significant associations within this group.

Phenotype	CNV Call Set	CNV Type	CNV Metric	No. of Obs.	z	P>|z|
Factor 2	Intersected	All	No. of CNVs	1924	-0.54	0.586
			CNV Burden (Kb)	1924	0.43	0.665
		Deletions	No. of CNVs	1924	-0.68	0.497
			CNV Burden (Kb)	1924	0.46	0.646
		Duplications	No. of CNVs	1924	0.65	0.515
			CNV Burden (Kb)	1924	0.55	0.585
	PennCNV	All	No. of CNVs	1924	0.09	0.931
			CNV Burden (Kb)	1924	0.54	0.587
		Deletions	No. of CNVs	1924	-0.58	0.560
			CNV Burden (Kb)	1924	0.49	0.621
		Duplications	No. of CNVs	1924	1.19	0.233
			CNV Burden (Kb)	1924	0.52	0.601
	iPattern	All	No. of CNVs	1924	1.20	0.230
			CNV Burden (Kb)	1924	1.08	0.280
		Deletions	No. of CNVs	1924	-0.58	0.564
			CNV Burden (Kb)	1924	0.48	0.631
		Duplications	No. of CNVs	1924	1.46	0.146
			CNV Burden (Kb)	1924	1.06	0.291
	QuantiSNP	All	No. of CNVs	1924	0.14	0.890
			CNV Burden (Kb)	1924	0.83	0.407
		Deletions	No. of CNVs	1924	0.28	0.781
			CNV Burden (Kb)	1924	0.94	0.346
		Duplications	No. of CNVs	1924	0.06	0.951
			CNV Burden (Kb)	1924	0.12	0.907

Table 6.4. Association analysis of CNV burden called with different methods with the guilt and psychomotor agitation symptoms dimension. Bonferroni adjusted p value for significance = 0.0021.

6.4.1.3.3 Factor 3 - Atypical Depressive Features

Fig. 6.5 illustrates the distribution of atypical depressive (increased appetite and hypersomnia) symptoms, as defined by dimensions of SCAN symptoms in Korszun et al.(Korszun et al., 2004).

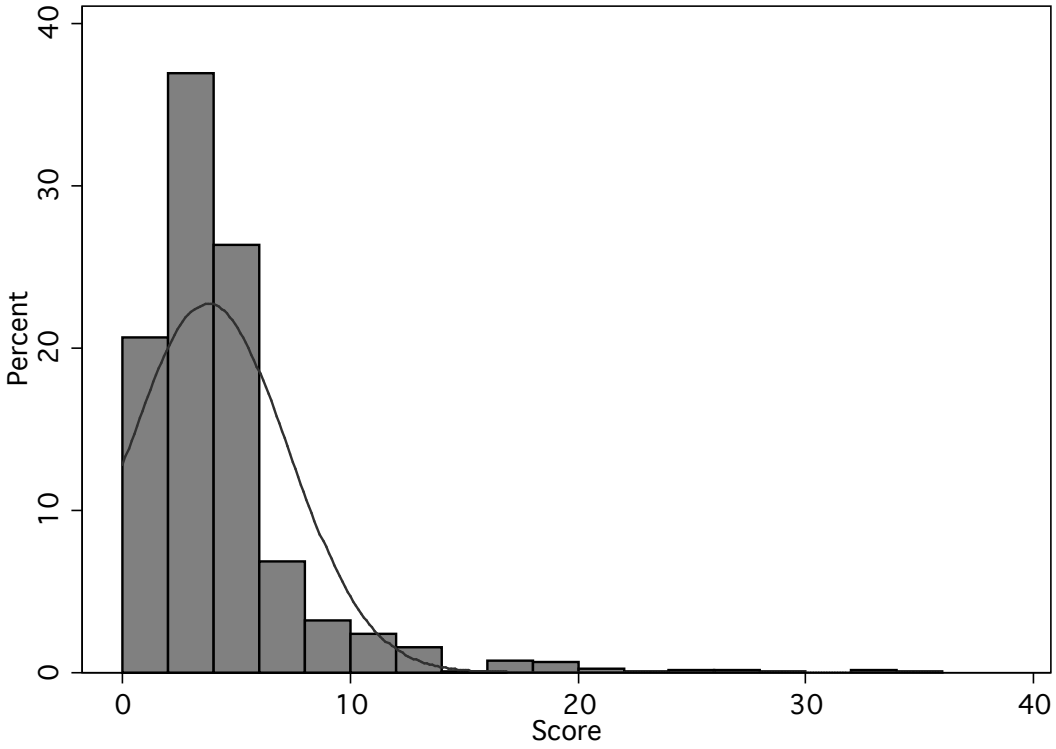


Fig. 6.5. Histogram of the atypical depressive (increased appetite and hypersomnia) symptoms dimension with a superimposed normal distribution plot.

Table 6.5 illustrates the results of bootstrapped linear regression analysis for the number of CNVs and total length of CNVs per person compared to the atypical depressive symptoms corrected for age, centre of recruitment, sex, and LRRSD. There are no significant associations within this group.

Phenotype	CNV Call Set	CNV Type	CNV Metric	No. of Obs.	z	P>|z|
Factor 3	Intersected	All	No. of CNVs	1924	-0.18	0.858
			CNV Burden (Kb)	1924	0.62	0.538
		Deletions	No. of CNVs	1924	-0.52	0.603
			CNV Burden (Kb)	1924	1.37	0.170
		Duplications	No. of CNVs	1924	0.49	0.622
			CNV Burden (Kb)	1924	-0.94	0.349
	PennCNV	All	No. of CNVs	1924	0.26	0.793
			CNV Burden (Kb)	1924	0.80	0.426
		Deletions	No. of CNVs	1924	-0.06	0.950
			CNV Burden (Kb)	1924	1.41	0.157
		Duplications	No. of CNVs	1924	0.51	0.612
			CNV Burden (Kb)	1924	-1.20	0.230
	iPattern	All	No. of CNVs	1924	-1.06	0.290
			CNV Burden (Kb)	1924	0.25	0.804
		Deletions	No. of CNVs	1924	-1.84	0.066
			CNV Burden (Kb)	1924	0.64	0.521
		Duplications	No. of CNVs	1924	-0.36	0.716
			CNV Burden (Kb)	1924	-1.27	0.203
	QuantiSNP	All	No. of CNVs	1924	-0.73	0.467
			CNV Burden (Kb)	1924	0.36	0.721
		Deletions	No. of CNVs	1924	-0.60	0.550
			CNV Burden (Kb)	1924	1.40	0.163
		Duplications	No. of CNVs	1924	-0.50	0.616
			CNV Burden (Kb)	1924	-1.92	0.054

Table 6.5. Association analysis of CNV burden called with different methods with the atypical depressive (increased appetite and hypersomnia) symptoms. Bonferroni adjusted p value for significance = 0.0021.

6.4.1.4 Personality Trait Analyses

6.4.1.4.1 Neuroticism

Fig. 6.6 illustrates the distribution of trait neuroticism scores derived from the EPQ.

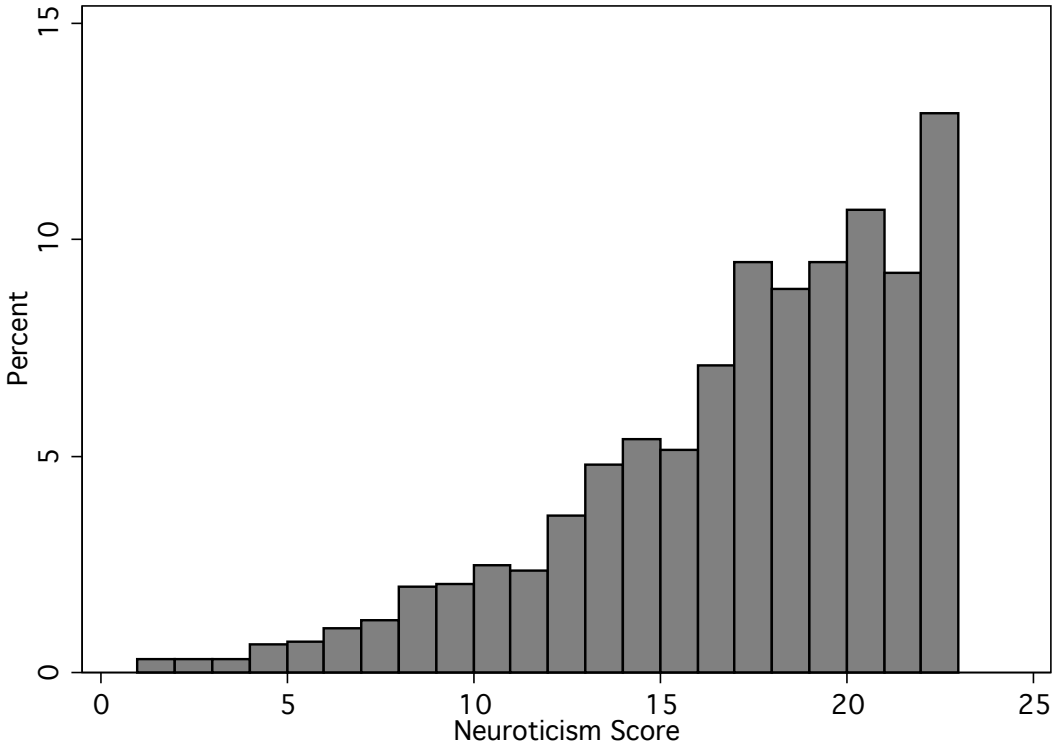


Fig. 6.6. Histogram of trait neuroticism scores.

Table 6.6 illustrates the results of bootstrapped linear regression analysis for the number of CNVs and total length of CNVs per person compared to trait neuroticism scores corrected for age, centre of recruitment, sex, and LRRSD. There are no significant associations within this group.

Phenotype	CNV Call Set	CNV Type	CNV Metric	No. of Obs.	z	P>|z|
Neuroticism	Intersected	All	No. of CNVs	1580	1.33	0.183
			CNV Burden (Kb)	1580	1.44	0.149
		Deletions	No. of CNVs	1580	2.10	0.035
			CNV Burden (Kb)	1580	2.10	0.036
		Duplications	No. of CNVs	1580	-0.09	0.928
			CNV Burden (Kb)	1580	0.11	0.910
	PennCNV	All	No. of CNVs	1580	1.54	0.124
			CNV Burden (Kb)	1580	1.60	0.110
		Deletions	No. of CNVs	1580	1.27	0.205
			CNV Burden (Kb)	1580	2.00	0.046
		Duplications	No. of CNVs	1580	0.79	0.432
			CNV Burden (Kb)	1580	0.46	0.646
	iPattern	All	No. of CNVs	1580	2.10	0.036
			CNV Burden (Kb)	1580	1.84	0.066
		Deletions	No. of CNVs	1580	2.18	0.029
			CNV Burden (Kb)	1580	2.69	0.007
		Duplications	No. of CNVs	1580	0.72	0.473
			CNV Burden (Kb)	1580	0.07	0.947
	QuantiSNP	All	No. of CNVs	1580	1.87	0.061
			CNV Burden (Kb)	1580	1.56	0.118
		Deletions	No. of CNVs	1580	1.19	0.236
			CNV Burden (Kb)	1580	1.64	0.101
		Duplications	No. of CNVs	1580	1.45	0.146
			CNV Burden (Kb)	1580	0.86	0.389

Table 6.6. Association analysis of CNV burden called with different methods with overall trait neuroticism scores. Bonferroni adjusted p value for significance = 0.0021.

6.4.1.4.2 Extraversion

Fig. 6.7 illustrates the distribution of overall trait extraversion scores derived from the EPQ.

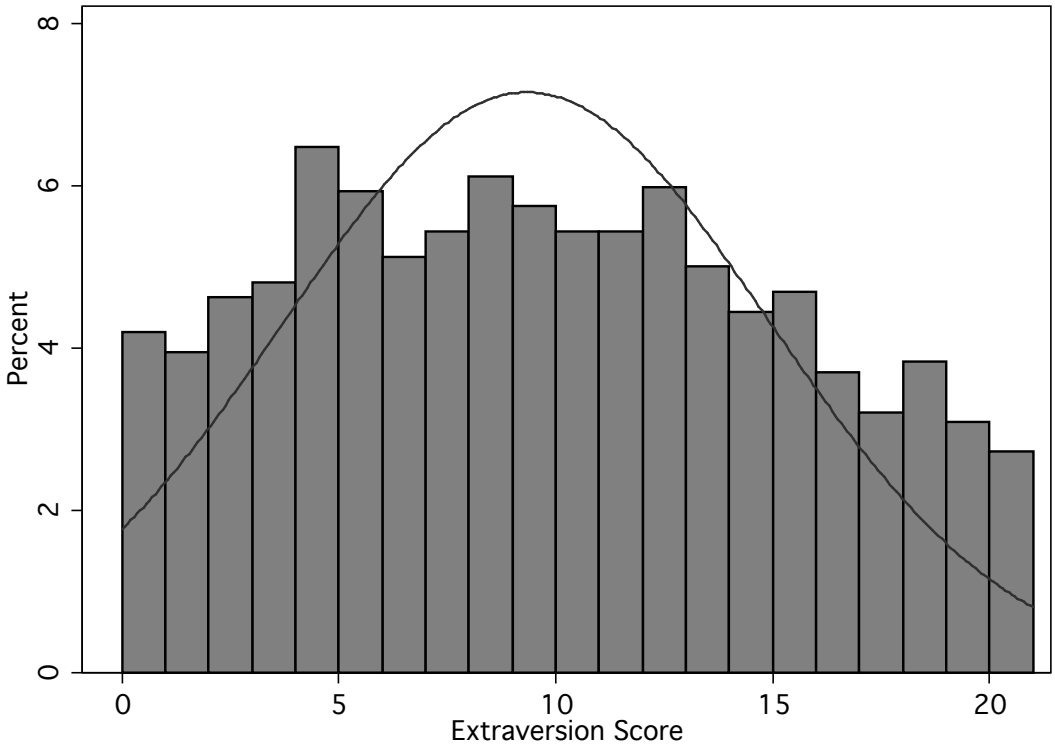


Fig. 6.7. Histogram of overall trait extraversion scores with a superimposed normal distribution plot.

Table 6.7 illustrates the results of bootstrapped linear regression analysis for the number of CNVs and total length of CNVs per person compared to overall trait extraversion scores corrected for age, centre of recruitment, sex, and LRRSD. There are no significant associations within this group.

Phenotype	CNV Call Set	CNV Type	CNV Metric	No. of Obs.	z	P>|z|
Extraversion	Intersected	All	No. of CNVs	1551	-0.16	0.874
			CNV Burden (Kb)	1551	-0.16	0.874
		Deletions	No. of CNVs	1551	-0.16	0.876
			CNV Burden (Kb)	1551	-0.16	0.873
		Duplications	No. of CNVs	1551	-0.16	0.870
			CNV Burden (Kb)	1551	-0.16	0.871
	PennCNV	All	No. of CNVs	1551	-0.16	0.875
			CNV Burden (Kb)	1551	-0.16	0.873
		Deletions	No. of CNVs	1551	-0.16	0.875
			CNV Burden (Kb)	1551	-0.16	0.875
		Duplications	No. of CNVs	1551	-0.16	0.871
			CNV Burden (Kb)	1551	-0.16	0.877
	iPattern	All	No. of CNVs	1551	-0.16	0.869
			CNV Burden (Kb)	1551	-0.16	0.871
		Deletions	No. of CNVs	1551	-0.16	0.872
			CNV Burden (Kb)	1551	-0.16	0.874
		Duplications	No. of CNVs	1551	-0.16	0.873
			CNV Burden (Kb)	1551	-0.16	0.873
	QuantiSNP	All	No. of CNVs	1551	-0.16	0.870
			CNV Burden (Kb)	1551	-0.16	0.872
		Deletions	No. of CNVs	1551	-0.17	0.869
			CNV Burden (Kb)	1551	-0.16	0.871
		Duplications	No. of CNVs	1551	-0.16	0.871
			CNV Burden (Kb)	1551	-0.16	0.871

Table 6.7. Association analysis of CNV burden called with different methods with overall trait extraversion scores. Bonferroni adjusted p value for significance = 0.0021.

6.4.1.4.3 Psychoticism

Fig. 6.8 illustrates the distribution of overall trait psychoticism scores derived from the EPQ.

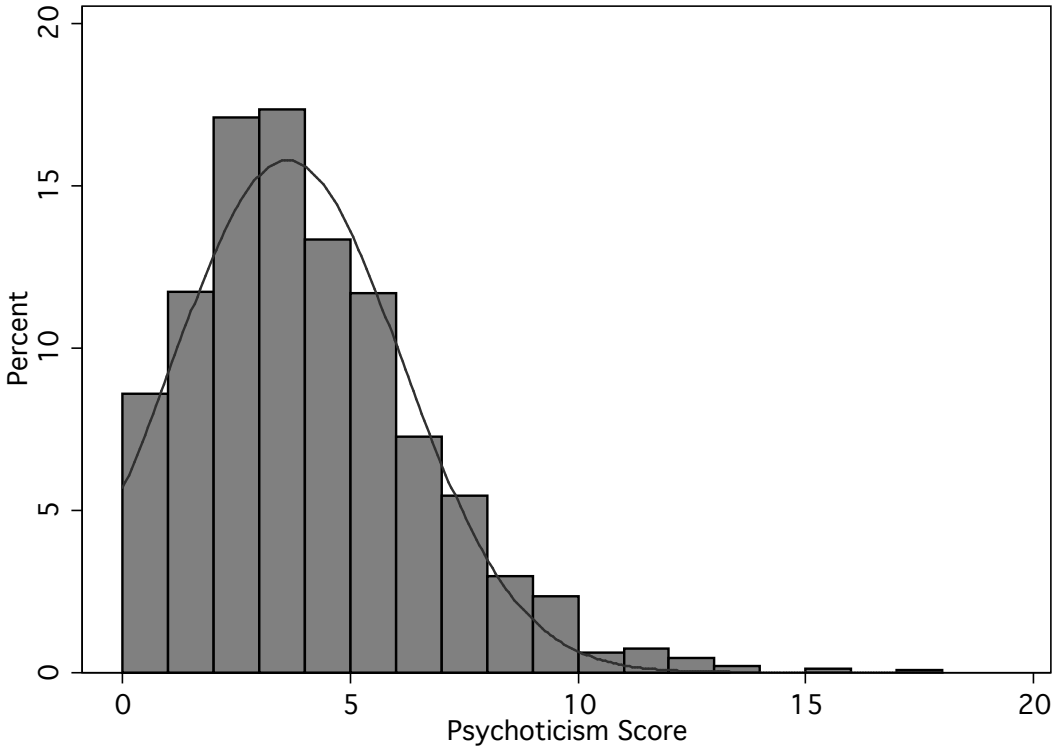


Fig. 6.8. Histogram of overall trait psychoticism scores with a superimposed normal distribution plot.

Table 6.8 illustrates the results of bootstrapped linear regression analysis for the number of CNVs and total length of CNVs per person compared to overall trait psychoticism scores corrected for age, centre of recruitment, sex, and LRRSD. There are no significant associations within this group.

Phenotype	CNV Call Set	CNV Type	CNV Metric	No. of Obs.	z	P>|z|
Psychoticism	Intersected	All	No. of CNVs	1480	0.33	0.740
			CNV Burden (Kb)	1480	0.22	0.828
		Deletions	No. of CNVs	1480	0.26	0.796
			CNV Burden (Kb)	1480	-0.04	0.964
		Duplications	No. of CNVs	1480	0.32	0.748
			CNV Burden (Kb)	1480	0.44	0.661
	PennCNV	All	No. of CNVs	1480	0.30	0.766
			CNV Burden (Kb)	1480	0.51	0.612
		Deletions	No. of CNVs	1480	0.07	0.947
			CNV Burden (Kb)	1480	0.09	0.929
		Duplications	No. of CNVs	1480	0.72	0.470
			CNV Burden (Kb)	1480	0.77	0.441
	iPattern	All	No. of CNVs	1480	0.88	0.380
			CNV Burden (Kb)	1480	0.51	0.611
		Deletions	No. of CNVs	1480	0.53	0.599
			CNV Burden (Kb)	1480	0.34	0.730
		Duplications	No. of CNVs	1480	0.51	0.613
			CNV Burden (Kb)	1480	0.19	0.851
	QuantiSNP	All	No. of CNVs	1480	0.87	0.386
			CNV Burden (Kb)	1480	1.18	0.239
		Deletions	No. of CNVs	1480	0.13	0.896
			CNV Burden (Kb)	1480	0.68	0.497
		Duplications	No. of CNVs	1480	0.98	0.326
			CNV Burden (Kb)	1480	0.91	0.365

Table 6.8. Association analysis of CNV burden called with different methods with overall trait psychoticism scores. Bonferroni adjusted p value for significance = 0.0021.

6.4.2 Sex Chromosome Abnormalities

We hypothesised that sex chromosome abnormalities might be more frequent in our dataset, based on previous research showing that those with sex chromosome abnormalities were at higher risk of psychiatric sequelae(Bender, Harmon, Linden, & Robinson, 1995; Ratcliffe SG et al., 1982).

To look for sex chromosome abnormalities we manually viewed plots of the LRR and BAF for the X and Y chromosome for all samples where the phenotypic sex did not match the sex inferred by the heterozygosity of the B allele frequency of X chromosome markers calculated by PennCNV. In practice, this is achieved by assigning a sex of female to any case where >10% of X chromosome markers have a B allele frequency of >0.25 & <0.75, which would usually denote a heterozygous allele call.

We compared the frequency of sex chromosome abnormalities in our case set with that detected in 34,910 sequentially screened liveborn infants in Denmark(Nielsen & Wohlert, 1991).

We detected 2 cases of 47,XXY (Klinefelter's syndrome) out of a total of 909 male cases (Fig. 6.9). We compared this frequency to that published in(Nielsen & Wohlert, 1991) (27 cases of Klinefelter's or mosaic Klinefelter's in 17,827 male liveborns) using a 1-sided Fisher's exact test for association. This was not significant (p=0.41, OR=1.46 (95%CI 0 - 5.54)).

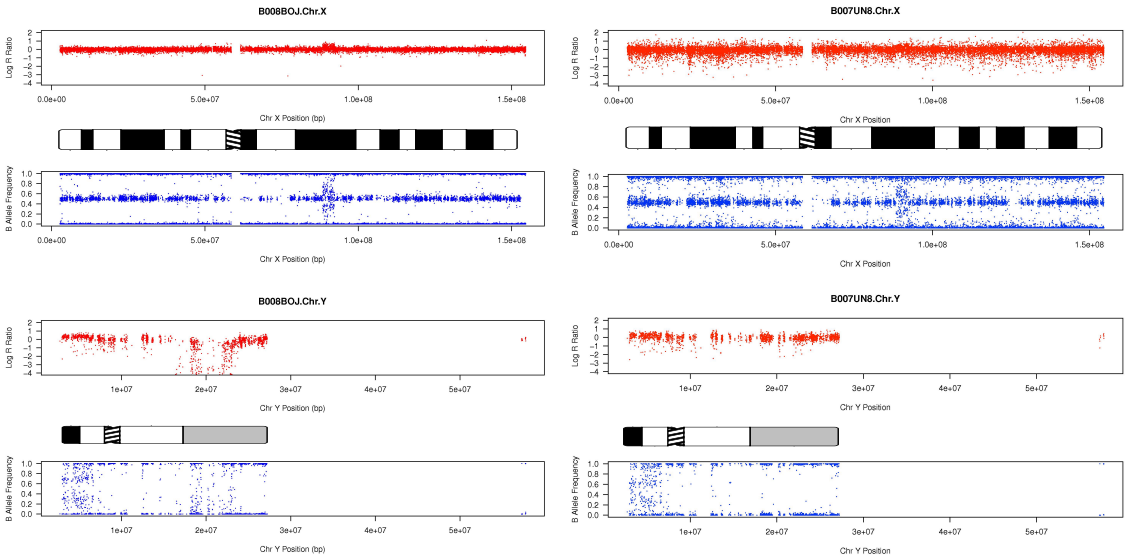


Fig. 6.9. 2 cases of Klinefelter's syndrome in our cases. Left panel - phenotypic male with evidence of two X chromosomes and a Y chromosome with a deletion of Yq. Right panel - phenotypic male with evidence of two X chromosomes and a Y chromosome.

Out of 2,197 female cases we detected 3 cases of 45,X (Turner's syndrome) (Fig. 610) of which 2 were 45,X/46,XX mosaics. We compared this frequency to that published in(Nielsen & Wohlert, 1991) (6 cases of Turner's or mosaic Turner's in 17,038 female liveborns) using a 1-sided Fisher's exact test for association. This was not significant (p=0.074, OR=3.88 (95%CI 1.06 - 14.17)).

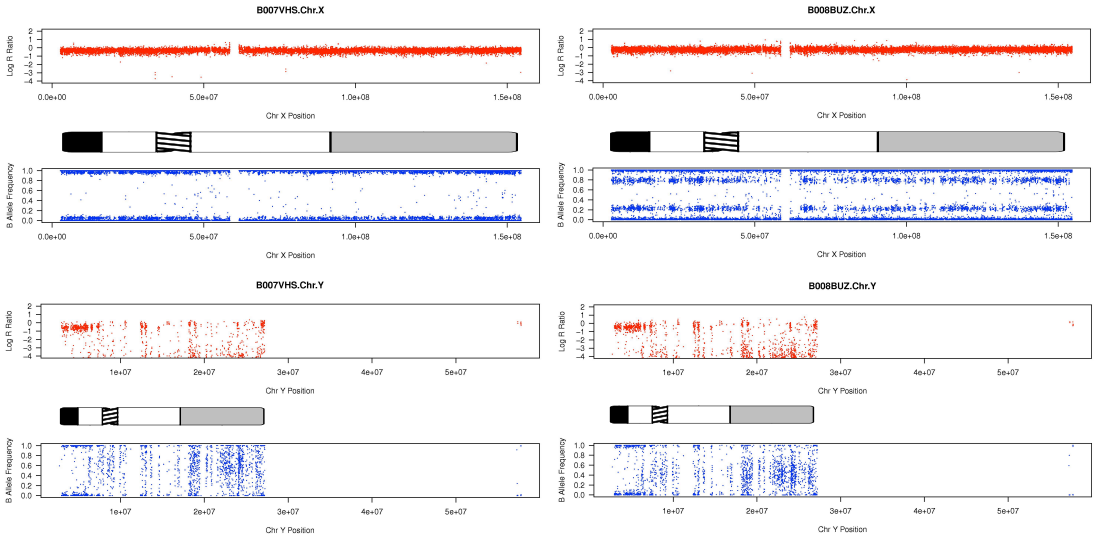


Fig. 6.10. 2 cases of Turner's syndrome in our cases. Left panel - phenotypic female with evidence of 1 X chromosome and no Y chromosome. Right panel - phenotypic female with evidence of X/XX mosaicism (split B allele frequency plot) and no Y chromosome.

6.4.3 Diploid/Triploid Mosaicism

We observed one case with plots that suggested a diploid/triploid mosaicism (Fig. 6.11.). This is, however, likely to be a somatic event as liveborns with true diploid/triploid mosaicisms confirmed by karyotype analysis usually have a severe phenotype comprising mental retardation, body asymmetry, syndactyly, facial abnormalities and a variety of other problems(van de Laar et al., 2002). Only 25 patients have ever been reported with such an abnormality in the literature. In the context of our study, therefore, this is most likely to represent an acquired mosaicism. The sample in question is derived from the venous blood of a 55 year old female. Little systematic research has been performed, however that which has been done suggests that somatic diploid/triploid mosaicisms are in fact relatively common, and may be a function of advancing age in cells derived from blood(Rodriguez-Santiago et al., 2010).

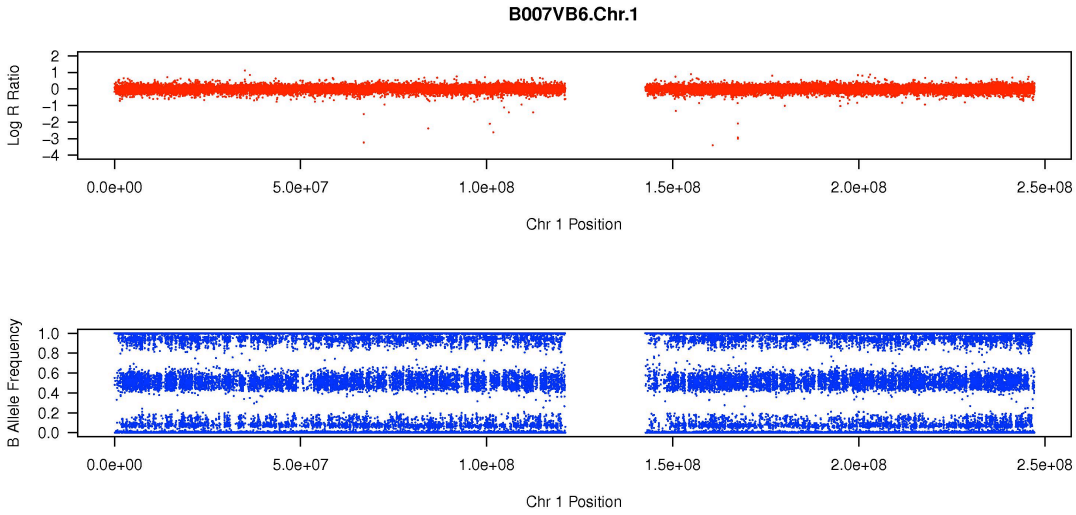


Fig. 6.11. Diploid/Triploid mosaicism in a case sample (lower plot- B allele frequency split six ways). Chromosome 1 plots shown.

6.4.4 In-Depth Analyses of Single Cases

Some cases were observed to have interesting CNVs, which we defined as those that were large, in regions previously highlighted by other studies in neuropsychiatric disorders or covering regions with genes pertinent to neuronal development. A small cohort of participants in the GWAS were later followed up with psychometric testing, and we also preferentially present data from these cases if they have interesting CNVs. On this basis, and in the final section of this chapter we report a series of cases with interesting CNVs, with clinical information presented as it was available.

6.4.4.1 3MB Deletion in 22q11.2

B007ULM at the time of entry to the GWA study was a 35 year old, right-handed, Caucasian male, employed, living with parents, single and with no children. His weight was 70kg and height 1.68m. He had a past medical history of

nephrectomy at the age of eight months but no more details were available. He was educated to degree level. He suffered his first depressive symptoms aged 14 and was first diagnosed with depressive disorder at 18. He was treated successfully with Clomipramine for 6 years from age 18-24, which led to remission of symptoms. Since then he has suffered a relapsing-remitting course of depression and a chronic course of moderate depressive symptoms for 4 years prior to entry into the GWA study, treated again with Clomipramine, a tricyclic antidepressant. There was no history of psychotic symptoms and no history of comorbid drug misuse. He had never been hospitalised for his depressive illness, had no history of treatment with electroconvulsive therapy, but had tried to commit suicide in the past. He had a past medical history of hypothryroidism, treated currently with thyroxine 175mcg daily. Blood workup at the time of entry to the GWAS showed normal levels of free thyroxine and thyroid stimulating hormone, a normal full blood count, normal renal function and normal liver function. He harboured a rare 3MB deletion of 22q11.2 that is associated with a variety of developmental and psychiatric abnormalities (Fig. 4.6, chapter 4.4.5).

At age 37 he returned for a follow up study. At this point he was still mildly depressed with a Beck Depression Inventory (BDI) score of 17. Full scale IQ measurement revealed a score of 101, with a verbal IQ of 104 and a performance IQ of 98. A neuroradiology report of the volumetric T1 weighted scan (blind to status) noted “a slight generalised volume loss in excess of that expected for age” (Dr. Sachit Shah, personal communication). No major intracranial abnormalities

were otherwise noted. Fig. 6.12. illustrates a sagittal, coronal and axial view of the T1 weighted scan.

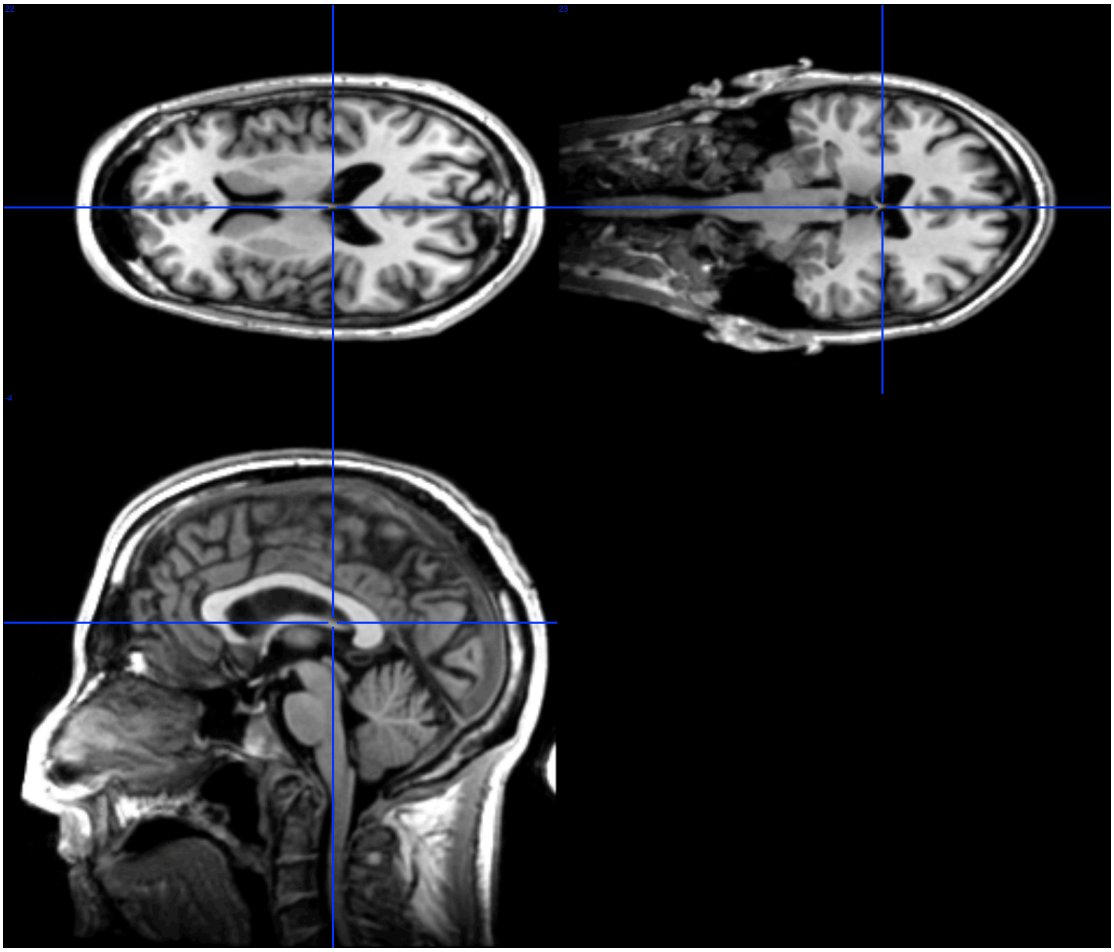


Fig. 6.12. T1 weighted volumetric MRI on B007ULM. No major intracranial abnormalities were noted, but there was a slight generalised volume loss in excess of that expected for age. No major craniofacial abnormalities were noted.

6.4.4.2 Singleton Duplication of the DISC1 Gene

B007V7N was found to have a 100kb duplication in chromosome 1, over exons 10 and 11 of the isoform variants L and Lv of the gene *DISC1* (disrupted in schizophrenia 1) (Fig. 6.13). *DISC1* has been extensively studied since a translocation disrupting this gene was found in a Scottish family with high

psychiatric morbidity(St Clair et al., 1990). The breakpoints in this family (exons 8,9) are close to, but not covered, by this CNV.

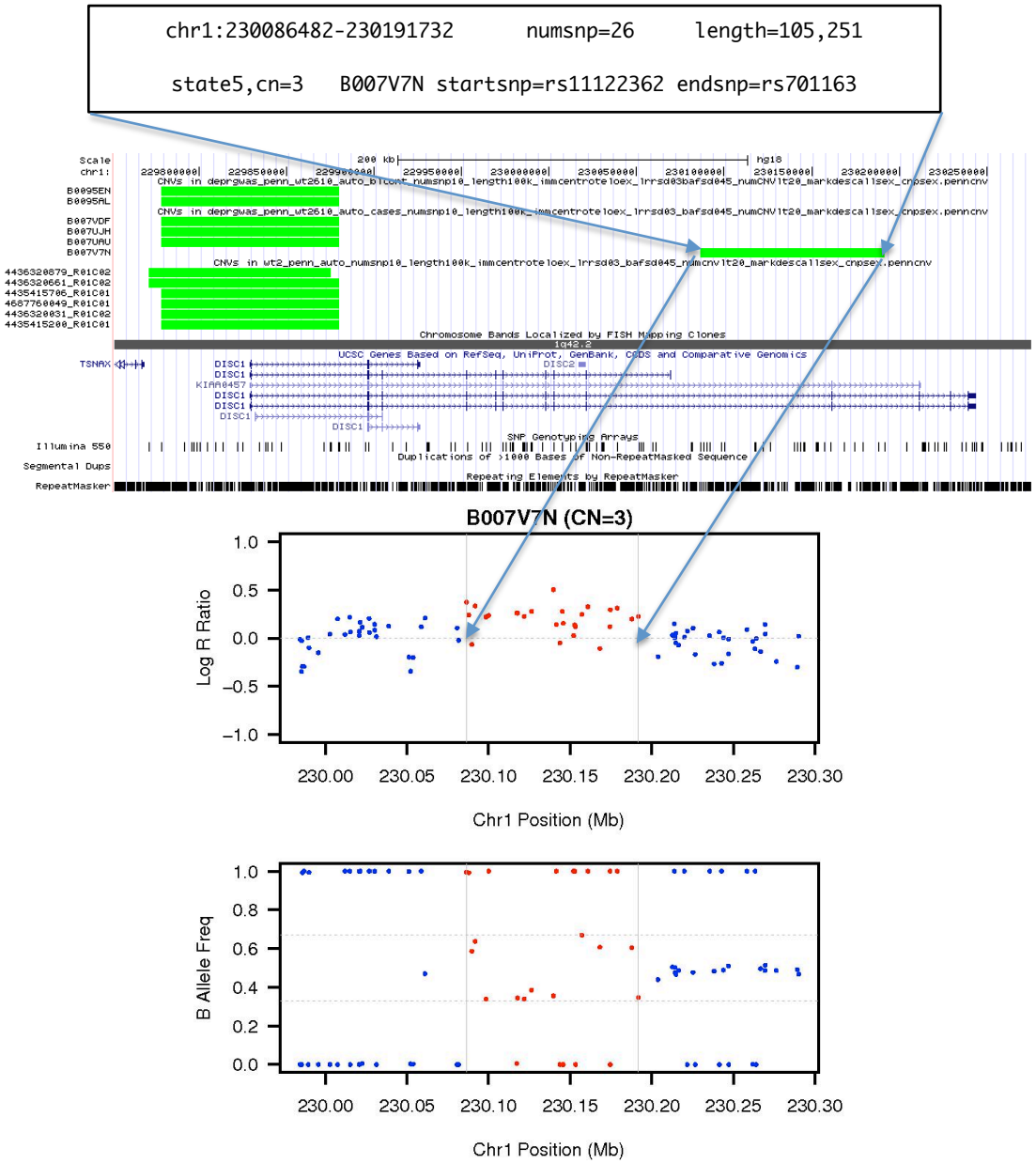


Fig. 6.13. B007V7N harbours a 100kb duplication over the isoform variants Es, L and Lv of the gene *DISC1* (Disrupted in schizophrenia 1).

B007V7N was, at the time of entry into the study, a 39 year old Caucasian female who suffered from recurrent, severe depressive episodes without psychotic

symptoms. At follow up 6 years later she had suffered three severe depressive episodes, at least 1 requiring hospitalisation and her current BDI score was 41 (severe depression). However she had no history of psychotic symptoms, suicide attempts, treatment with electroconvulsive therapy and no history of comorbid medical disorders or drug and alcohol misuse. She had a full scale IQ of 104 with a significant discrepancy (>20 points) between her verbal IQ of 94 and her performance IQ of 117.

6.4.5.3 Deletion of the GPC5 Gene

B007URQ was, at the time of entry to the GWAS, a 38 year old year old Caucasian lady, married with two children. Her mother had suffered from post-natal depression after her birth, and her father had suffered from chronic depression all her life until he died when B007URQ was 17 years old. Subsequent to this she suffered from anorexia for 2 years. She, however, achieved A-levels and a postgraduate degree. She had no past medical history of note and no history of drug or alcohol misuse or psychosis. She developed her first depressive episode in 2003 following a change of career. She was only partially responsive to multiple therapies, tried to commit suicide on numerous occasions and was twice treated with electroconvulsive therapy. At follow up 2 years later she was found to have a full scale IQ of 120 with a significant discrepancy between her verbal IQ (129) and performance IQ (108). Her BDI at follow up was 40 (severe depression).

B007URQ was found to have a rare deletion of the gene *GPC5* (Glypican 5) (Fig. 6.14). This deletion was also seen in three WTCCC2 controls. A run of

homozygous allele calls flank the deletion in this case. GPC5 has not been associated with any specific disorders, however it is interesting because it is highly expressed in the developing CNS, and particularly in the adult caudate, amygdala and prefrontal cortex(Watanabe, Yamada, & Yamaguchi, 1995).

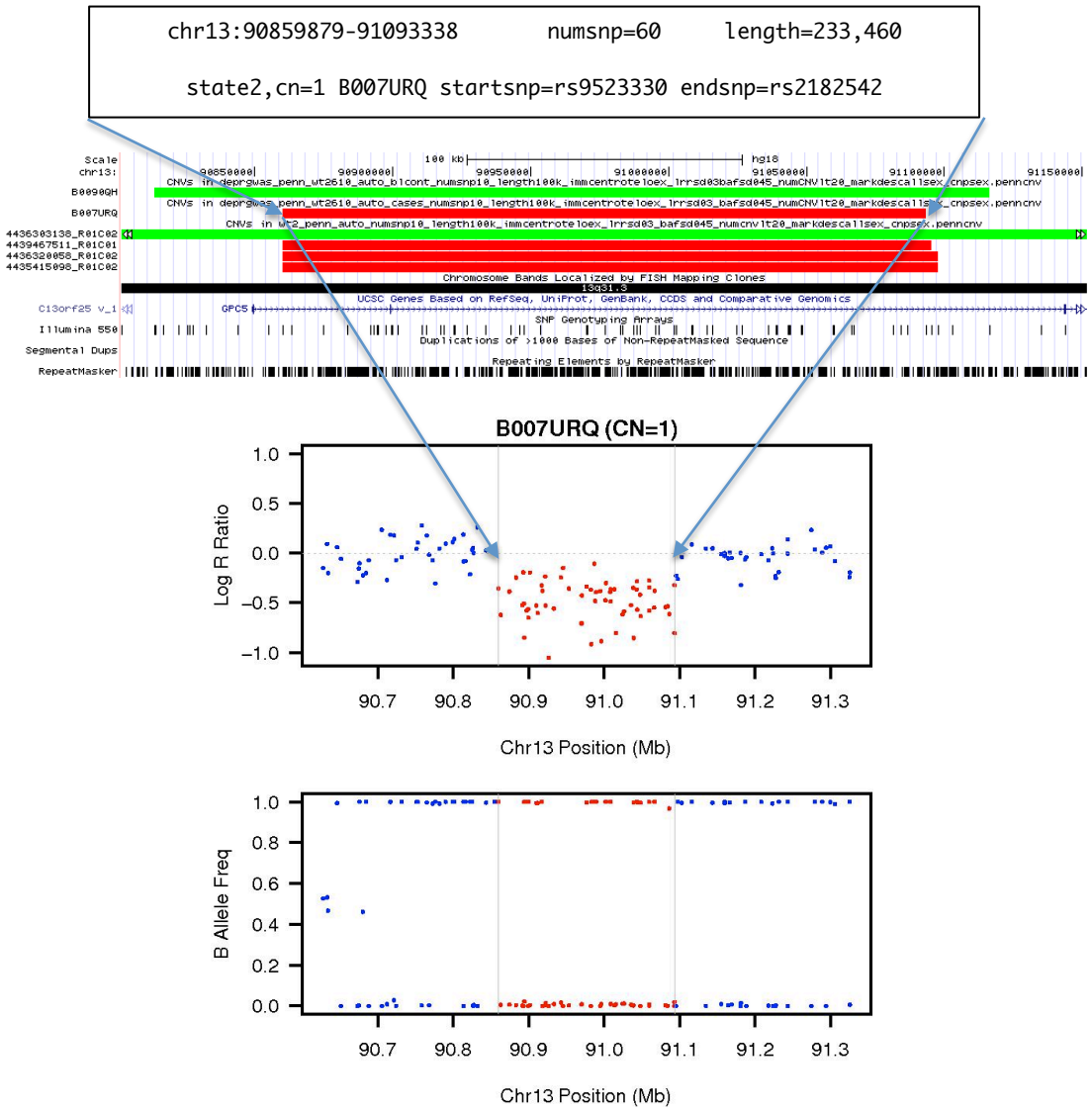


Fig. 6.14. B007URQ harbours a rare deletion CNV affecting the gene *GPC5* (Glypican 5).

A run of homozygous allele calls flank the deletion in this case (lower panel, B Allele Freq plot).

6.4.4.4 Singleton Deletion of 9MB in Chromosome 13

The largest CNV in our case cohort was a 9MB deletion seen over a relatively gene-sparse region of chromosome 13 in a female case who had suffered 2 severe depressive episodes but otherwise had an unremarkable past medical history (Fig. 6.15). The proximal breakpoint of the variant interrupts isoform 1 of the gene *PCDH9 (*protocadherin 9) which is ubiquitously and highly expressed in brain, but particularly in the prefrontal cortex.

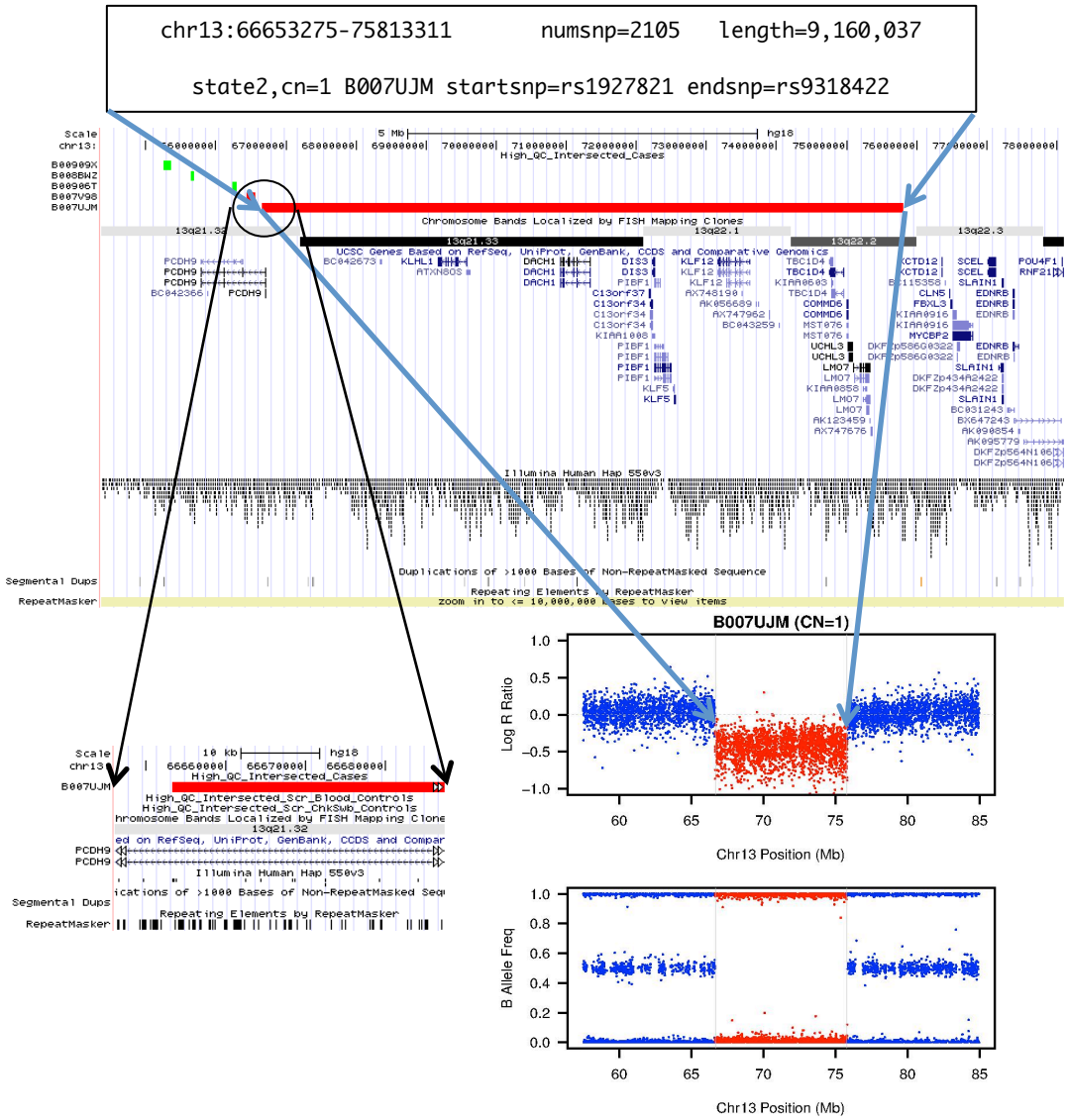


Fig. 6.15. B007UJM harbours a 9MB deletion CNV interrupting *PCDH9 (*protocadherin 9) (lower left panel) and deleting one copy of several other genes.

6.4.4.5 Singleton Duplication of the CACNA1C Gene

One case sample has a singleton duplication event over and interrupting *CACNA1C* (calcium channel, voltage-dependent, L type) (Fig. 6.16), which has robust evidence of association with mood disorder, recently implicated in a meta-analysis in bipolar disorder and schizophrenia by the psychiatric GWAS consortium(Ripke et al., 2011). This case was confirmed to have no history of bipolar symptoms and no psychotic symptoms, but was found to have a very high trait neuroticism score (22, 4th quartile) and an early age of onset (14 years).

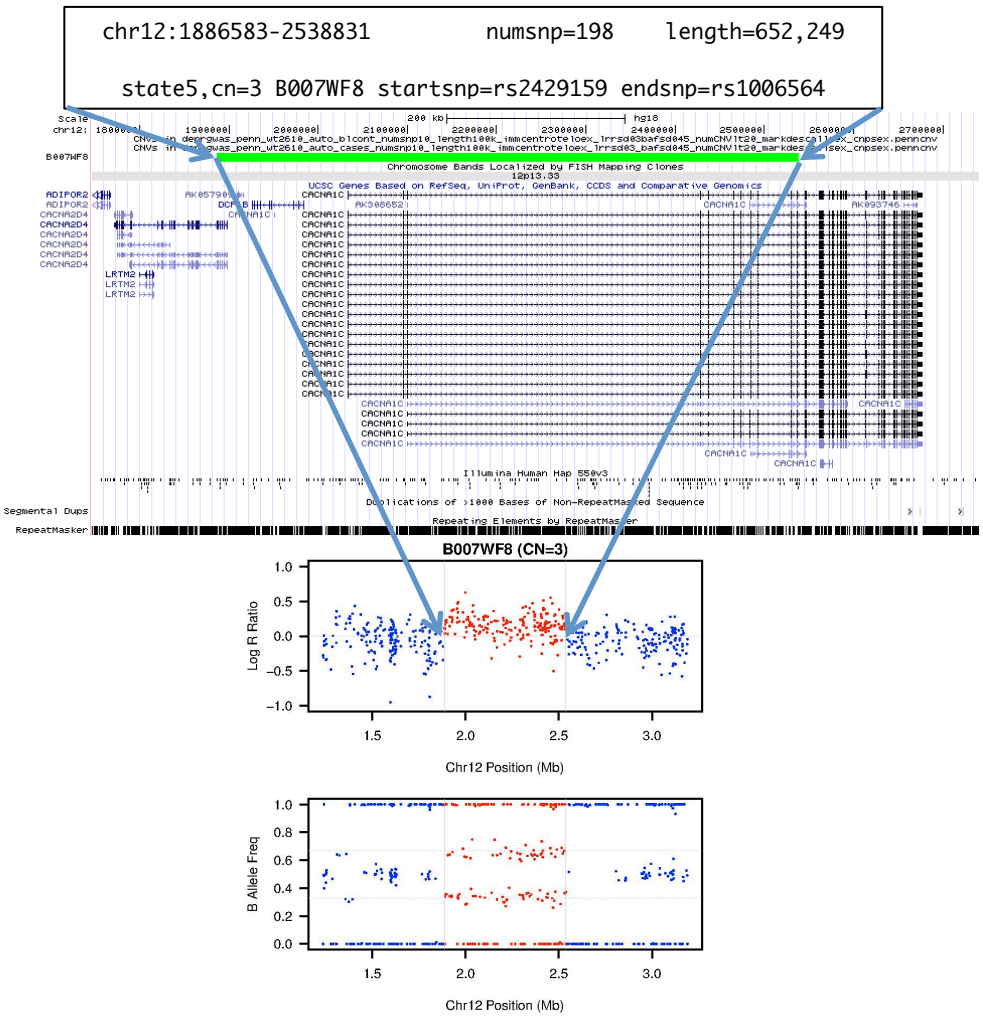


Fig. 6.16. B007WF8 harbours a singleton duplication CNV interrupting the gene CACNA1C.

6.4.4.6 Singleton Deletion of the Choline Transporter Gene

Another case harboured a large rare singleton deletion in chromosome 2 (Fig. 6.17), which, amongst other genes, deleted *SLC5A7 (*solute carrier family 5 (choline transporter)), which enables the uptake of choline into cholinergic neurones for the synthesis of acetylcholine.

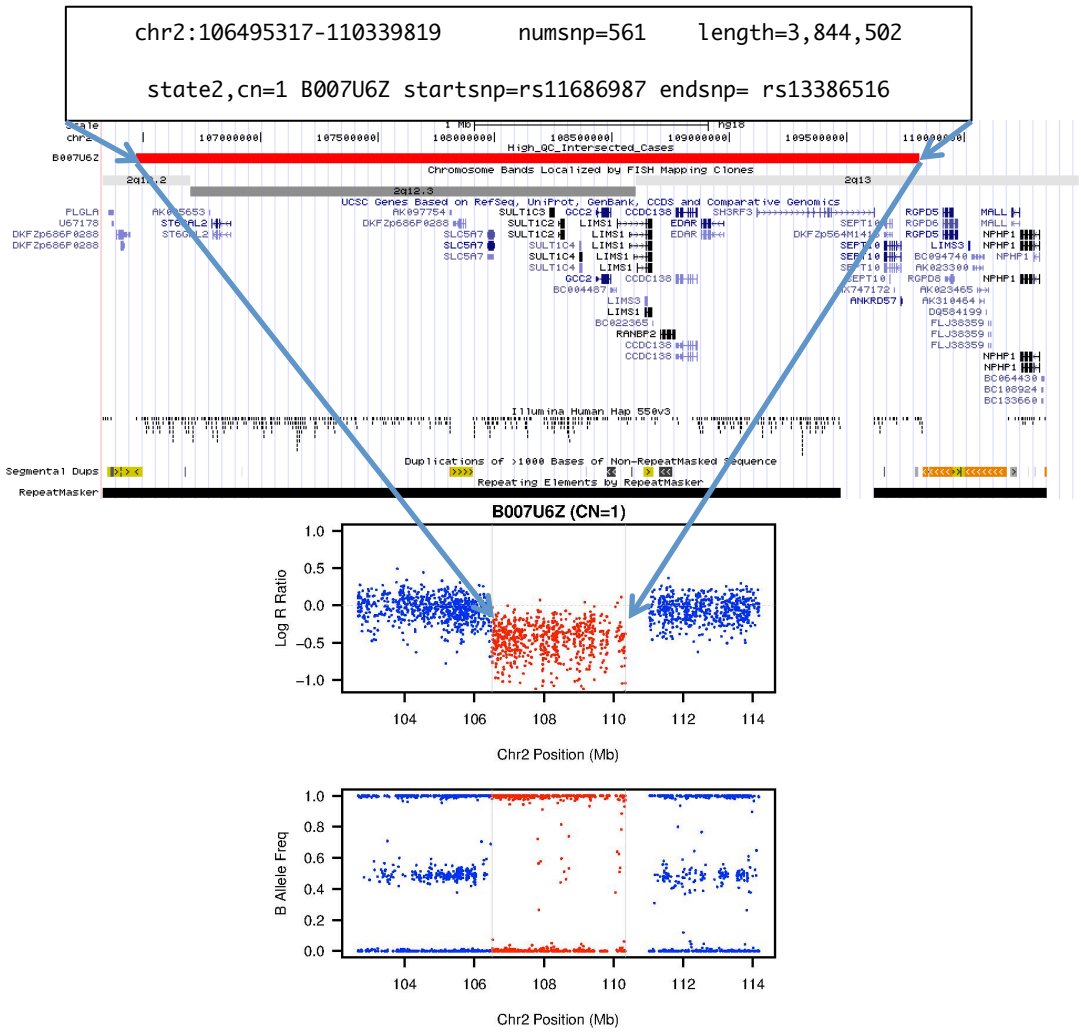


Fig. 6.17. B007U6Z harbours a large deletion CNV in chromosome 2, which deletes one copy of the gene *SLC5A7 (*solute carrier family 5 (choline transporter)), amongst other genes.

6.4.4.7 Singleton Duplication of the NRG3, PCDH21 and GRID1 Genes

Another case, with an age of onset of depressive disorder of 16 years and two severe depressive episodes without psychotic symptoms, harboured a large singleton duplication that interrupts the first exon of the *NRG3 (*neuregulin 3) gene, and duplicates the *PCDH21* (protocadherin 21) and the GRID1 (glutamate receptor, ionotropic, delta 1) gene (Fig. 6.18). All these genes are highly expressed in brain.

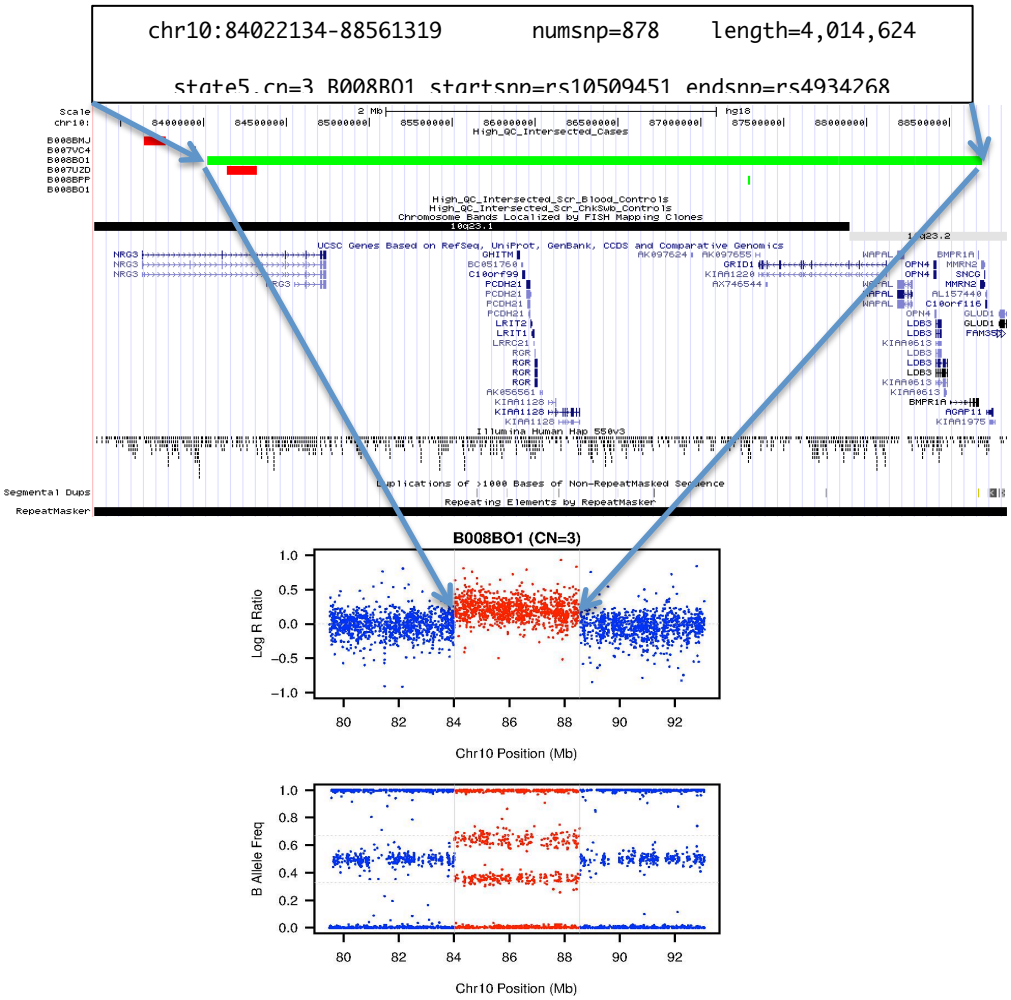


Fig. 6.18. B008BO1 harbours a large duplication CNV in chromosome 10 interrupting and spanning several genes highly associated with neuronal development and functioning.

6.4.4.8 Singleton Deletion of the GRIA4 Gene

B007VEM was a 32 year old woman who had suffered severe depressive episodes without psychotic symptoms from the age of 16. She exhibited the highest possible trait neuroticism score (23) and a high burden of depressive symptoms, with a BDI at interview of 41 (severe depression). She was found to have a large singleton deletion CNV in chromosome 11 deleting, amongst other genes, *GRIA4* (glutamate receptor, ionotrophic, AMPA 4 isoform) (Fig. 6.19). This CNV also deletes the *GUCY1A2 (*guanylate cyclase 1, soluble, alpha 2) gene and the proximal breakpoint interrupts the *PDGFD* (platelet derived growth factor D isoform 2) gene.

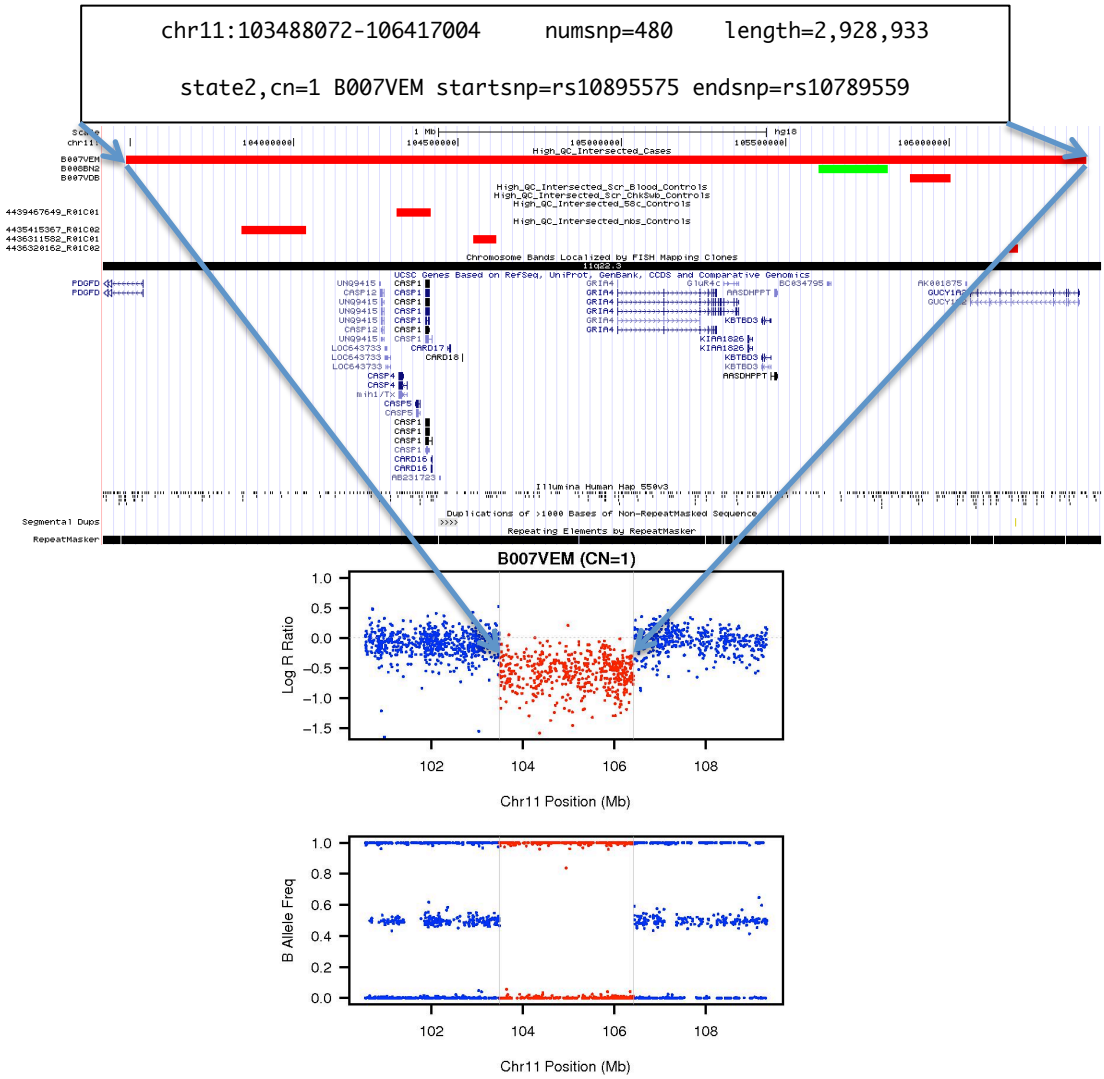


Fig. 6.19. B007VEM harbours a large, singleton deletion CNV affecting a number of genes, including a subunit of the AMPA sensitive glutamate receptor.

6.4.4.9 Singleton Deletion of the SLC6A15 Gene

B007WFG was a 25 year old woman at the time of interview and had suffered recurrent, severe depressive episodes since the age of 14, but without psychotic symptoms. She was found to carry a rare deletion CNV on chromosome 12 (Fig. 6.20) deleting, amongst others, the genes *SLC6A15 (*solute carrier family 6, member 15 isoform 1), *NTN (*neurotensin/neuromedin N preprotein) and interrupting MGAT4C (UDP-N-acetylglucosamine: alpha-1,3-D-mannoside).

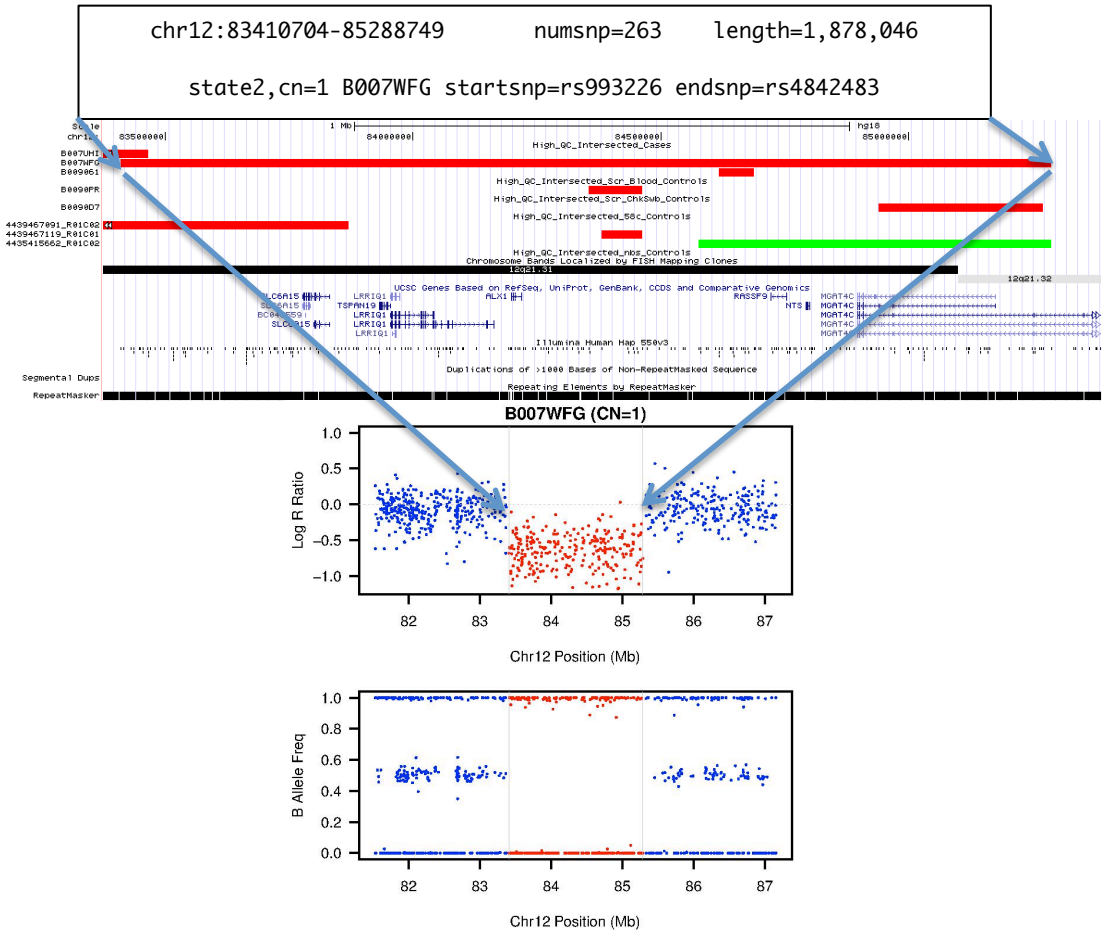


Fig. 6.20. B007WFG harbours a large singleton deletion CNV in chromosome 12 affecting, amongst others, the gene SLCA15.

6.4.4.10 Singleton Deletion Interrupting ZNF385D

B007UXK, at the time of entry into the GWAS, was a 52 year old Caucasian lady, married with three children. She has a past medical history of migraines and irritable bowel syndrome but there was no history of drug or alcohol misuse or psychosis. She had suffered recurrent depressive episodes since the age of 14. She was treated with a number of talking therapies and pharmacological treatments with partial remission of her symptoms. Both her mother and maternal grandfather had suffered from depression, with her grandfather also suffering manic episodes.

B007UXK was found to harbour a rare deletion interrupting the second exon of the zinc finger domain containing gene ZNF385D (Fig. 6.21). This deletion was not found in any other case or control sample (a singleton deletion). ZNF385D is a gene of unknown function, however transcriptional array studies suggest that it is relatively highly expressed in adult prefrontal cortex and cingulate cortex, parietal lobe and temporal lobe. It is also highly expressed in skeletal muscle, the foetal thyroid and the testes.

At follow up B007UXK was found to have a full-scale IQ of 124 with a verbal IQ of 129 and a performance IQ of 112. She remained depressed with a BDI score of 20 (moderate depression).

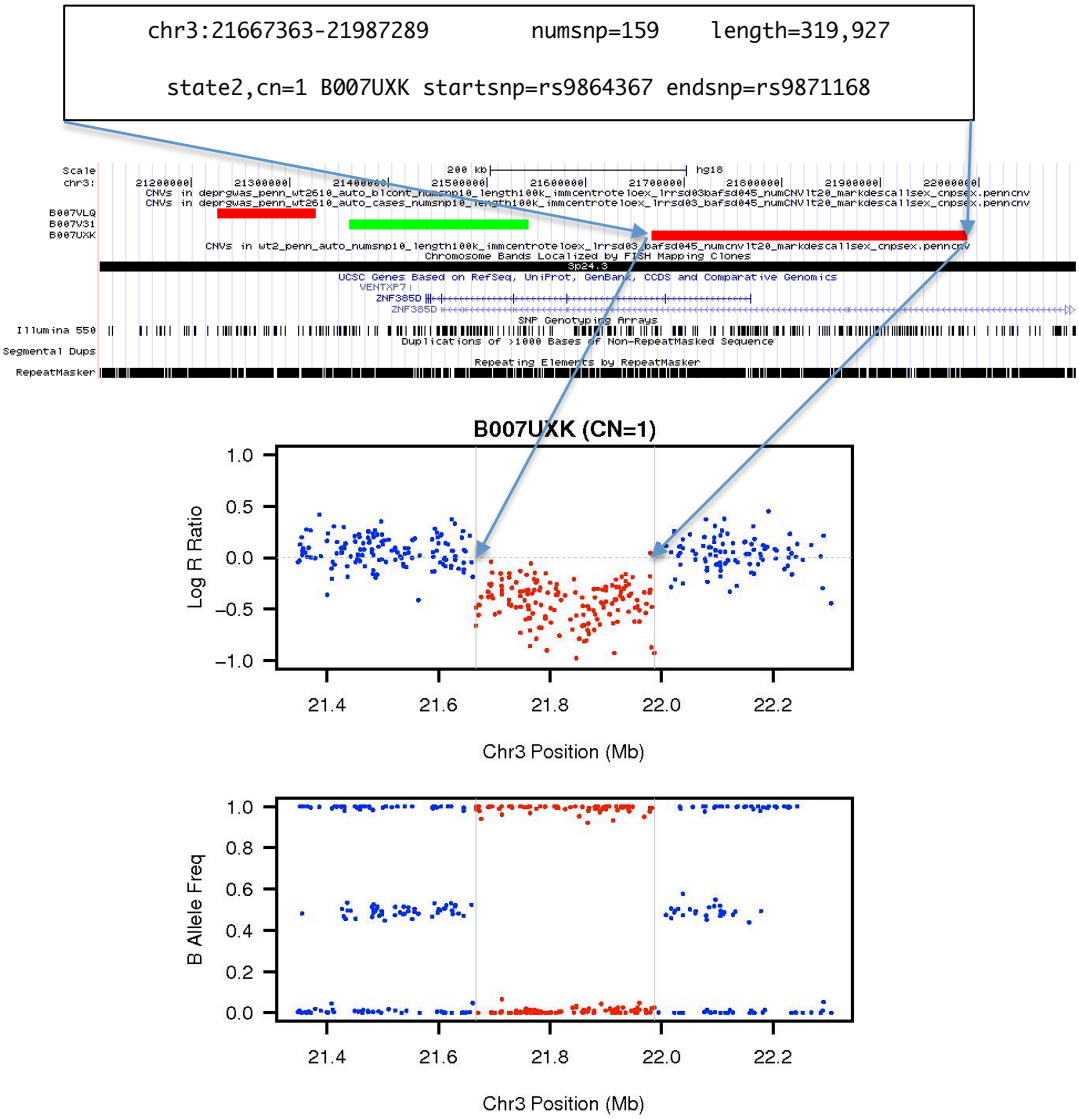


Fig. 6.21. B007UXK has a rare, singleton deletion CNV interrupting the ZNF385D gene, which is highly expressed in the cingulate and prefrontal cortex.

6.4.4.11 CNVs of the ABPA2 Gene

Two cases and two cases respectively have deletions and duplications in the 15q13.3 region, covering the *APBA2* (amyloid beta A4 precursor protein-binding) gene (Fig. 6.18). One deletion is also seen within the WTCCC2 controls (which may include individuals with psychiatric illness). B008BW9 (duplication) was a 44 year old woman at the time of entry into the study with a 25 year history of recurrent severe depressive episodes. It was unknown whether or not she had ever suffered psychotic symptoms. She had a moderate trait neuroticism score (14). B008BO7 (deletion) was a 49 year old woman at entry into the study. She reported a lifetime history of chronic depressive episodes, without any history of psychotic symptoms, since early teenage years. Her trait neuroticism score was 19 (3rd quantile). B007V4K (duplication) was a 55 year old lady at entry into the study, married with two children, with a history of recurrent depressive episodes of moderate severity and panic disorder since the age of 14. She had three healthy siblings with no history of depressive disorder. She had a trait neuroticism score of 17 (2nd quantile). B00908J (deletion) was a 22 year old lady who had suffered severe and recurrent depressive episodes since the age of 17. She had the highest possible trait neuroticism score (23). Unfortunately no further clinical information was available for these cases. Comparing the frequency of this CNV in cases (4 in 2,723) vs. WTCCC2 controls (1 in 4,828) with Fisher's exact test result in a borderline significant association (p=0.060, OR=7.1 (95%CI 1.07-NaN).

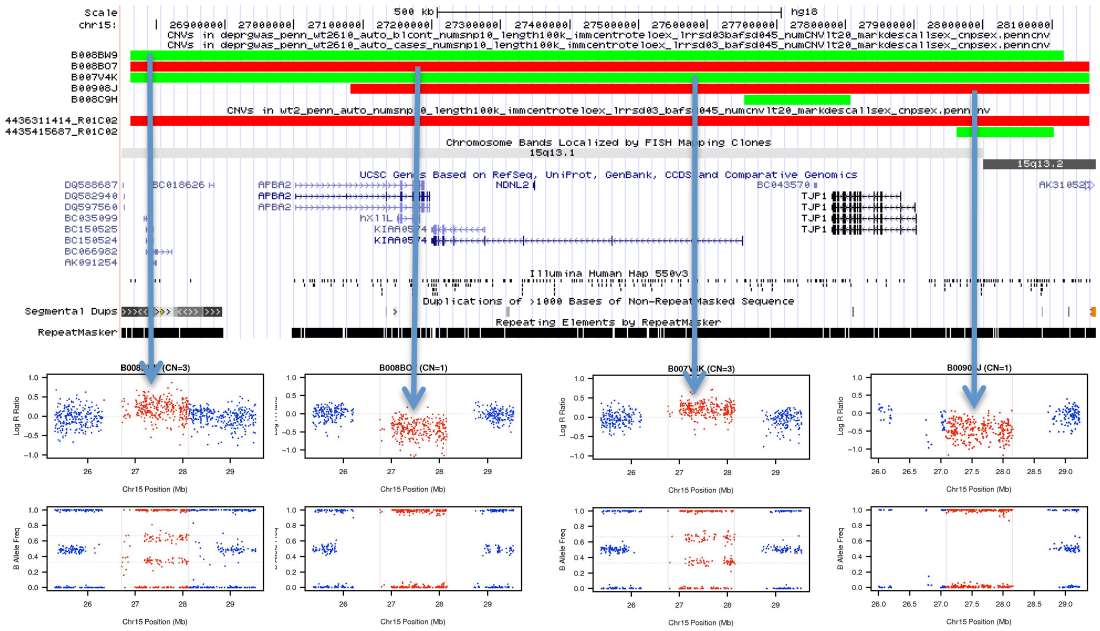


Fig. 6.22. Four cases have CNVs affecting a locus previously associated with mental retardation, schizophrenia and autism, and including the gene ABPA2. The fourth case (bottom panel, far right), on visual inspection of the plot, has a deletion CNV spanning the entire region rather than the partial region indicated in the upper panel.

These cases will be discussed further in the next chapter.

6.5 Conclusion

Phenotypic analysis comparing CNV burden with different phenotypes derived from a consensus phenotypic dataset derived from all cohorts within the GWA study did not show any significant differences after correction for multiple testing. We analysed the frequency of sex chromosome abnormalities in our cases compared to that seen in a large series of live born neonates, and failed to find a significant difference between the frequency of Klinefelter's syndrome and Turner's syndrome. Finally we illustrated a series of cases with interesting and putatively functional CNVs, some of whom we had collected additional phenotypic data on.

Chapter 7. Discussion

7.1 Introduction

In this final chapter I will first summarise our findings by chapter. I will then discuss our methodology and the limitations of our analysis methods. Next, I will discuss our findings and their relevance to the field of psychiatric genetics, and briefly discuss the field of medical genetics in complex diseases. Finally we will discuss how this research may be taken forward.

7.1.1 Summary of Cohorts and Analysis Methods by Chapter

A concise summary of the different analyses and cohorts used in this thesis is now given, along with the relevant chapter heading for further reference.

Chapter	Subsection Reference	Description of Cohorts and Analysis
Chapter 2. Rare CNVs. PennCNV.	2.4.1	RDD cases vs. screened controls and WTCCC2 controls. Fisher's exact tests of counts of samples with rare CNVs over whole genome/genic/exonic/ intronic/non-genic regions
	2.4.3	Identical to 2.4.1, except 10% of worst performing samples from 2.4.1 removed.
	2.4.4	Identical to 2.4.1. except UK only cases vs. UK only controls.
	2.4.5	Regions previously associated with schizophrenia. Cohorts identical to 2.4.1
	2.4.7	Cohorts identical to 2.4.1. Burden analyses using PLINK.
	2.4.8	Cohorts identical to 2.4.1. Singleton burden CNV analysis using PLINK.
Chapter 3. Common CNVs	3.4.1	UK RDD cases vs. UK screened controls derived from Lewis et al. (2010). GWA analysis of tagSNPs using logistic regression and permutation with PLINK.
	3.4.2	UK RDD cases vs. WTCCC2 population controls processed in identical manner to Lewis et al., (2010). GWA analysis identical to 3.4.1.
Chapter 4. Chromosome 22q11.2	4.4.1	RDD cases vs. screened controls over 22q11.2 region only using all 610 quad probes. Analysis with PLINK..
	4.4.3	Subset of RDD cases vs. screened controls genotyped with array CGH, called with DNA copy. Analysis with PLINK.
Chapter 5. iPattern and QuantiSNP methods	5.4.1.1	RDD cases vs. screened controls and WTCCC2 controls. Identical cohort and sample QC to 2.4.1. Three calling methods. Analysis with PLINK.
	5.4.1.2	Identical to 5.4.1.1 except high QC sample set described in 5.3.1.4.
	5.4.4	Identical to 2.4.5 except using high QC sample set and intersected calls described in 5.3.1.4 and 5.3.1.3 respectively.
Chapter 6. Phenotype Analyses.	6.4.1	Phenotypic association analysis with high QC RDD cases only as described in 5.3.1.4.

Table 7.1. Summary of Analysis Methods and Cohorts Used Throughout This Thesis.

7.2 Summary and Discussion of Results by Chapter

7.2.1 Rare CNVs

In chapter 2 we performed a case control association study of rare copy number variants. We hypothesised that, given the association of rare CNVs in other psychiatric disorders, that we might find a similar enrichment in our recurrent depression cohort. We compared cases of recurrent depression to a screened control cohort, and a population control cohort derived from phase 2 of the Wellcome Trust Case Control Consortium (WTCCC2). All samples were genotyped on Illumina Infinium HD arrays, but differed in the specific array type, altering slightly the number of markers. We therefore called CNVs with the PennCNV method using data from a consensus set of markers common to both array types used and processed our data to exclude poor quality samples and false positive calls.

In order to statistically analyse our data, we reduced our call set to a binary CNV present/CNV absent dataset, and then compared the proportion of samples with rare CNVs in each cohort, stratifying our calls into deletions and duplications and further stratifying by whether the CNVs covered exonic, intronic, genic or non-genic regions of the genome, as defined by RefSeq(Pruitt et al., 2009). The advantage of such an approach is that it allows the calculation of odds ratios to assess the strength of any observed association.

When the proportion of cases with a rare CNV was compared with the proportion of screened controls with a rare CNV, we found a striking difference,

with an almost 7% difference between the two cohorts (p=0.019 OR=1.31 (95%CI 1.04-1.64)). This was driven in the main by an almost 10% difference between the proportion of cases and screened controls to harbour a rare deletion CNV ($p=5.57x10^{-4}$ OR=1.52 (95%CI 1.20-1.93)). In contrast, no significant difference between the proportion of samples with a duplication was found (p=0.85, OR=1.02 (95%CI 0.81-1.28)). This was an interesting finding. The relatively consistent frequency of duplications across cohorts reassured us that systematic bias between our cohorts was less likely. Secondly it seemed plausible that recurrent depression cases might be enriched in deletion variants, and that this might confer a lack of cognitive resilience against depressive disorder.

It is reasonable to hypothesise that if deletion burden confers a lack of resilience against RDD, then it might occur over genic areas, and especially those genic areas encoding exons. We found that the degree of association with deletion CNVs increased step-wise in genic ($p=4.24x10^{-5}$, OR=1.78 (95% CI 1.35 - 2.35)) and exonic regions ($p=1.70x10^{-5}$, OR=1.87 (95% CI 1.40 - 2.49), but in non-genic regions there was no difference between cases and screened controls (p=0.31, OR=1.16 (95% CI 0.87 - 1.56). Again, the proportion of samples with duplications remained equivalent between the two groups across different regions of the genome.

We next tested for the same association in the cases and WTCCC2 controls. We again observed a similar difference, but not to the same magnitude, as reflected in a lower odds ratio, with our screened controls. The larger sample number

resulted, however, in a higher significance value. Cases were more likely to have a deletion than controls ($p=7.7x10^{-6}$, OR=1.25 (95% CI 1.13 - 1.37)) although in this instance we saw that the degree of association remained static in genic ($p=7.2x10^{-6}$, OR=1.27 (95% CI 1.14 -1.41)) and exonic regions ($p=1.04x10^{-5}$, OR=1.27 (95% CI 1.14 - 1.41)). Taken with the last observation made with the screened controls, our results showed that samples with deletions were least common in screened controls, middling in the WTCCC2 controls, and highest in our cases. If cases of RDD are enriched in numerous deletion CNVs that confer a lack of resilience we would expect to see a proportion of these in a population cohort, but at a lower overall frequency than in the RDD group.

We were concerned that our association could be being driven by false positive calls in low quality samples. To investigate this we removed the worst performing 10% of samples across our cohorts and reanalysed our data. We observed that the difference in frequency of samples with a rare deletion fell from 9.6% to 7.5% (p=0.017, OR=1.41 (95% CI 1.06 - 1.87), OR difference=-0.11) when cases were compared to screened controls, and from 5.2% to 3.5% (p=0.0055, OR=1.16 (95% CI 1.05 - 1.30), OR difference=-0.11) when cases were compared to WTCCC2 controls. Across exonic regions we observed that the difference in frequency fell from 10.9% to 7.5% (p=0.0060, OR=1.63 (95% CI 1.15 - 2.31), OR difference=-0.24) and from 4.6% to 2.2% (p=0.043, OR=1.14 (95% CI 1.00 - 1.29), OR difference=-0.13) when cases were compared to screened controls and WTCCC2 controls respectively. Whilst some reduction in significance is attributable to a reduction in power to detect an effect, the reduction in odds ratios suggests we also removed a proportion of samples at the

lower end of our QC range that were more likely to have contributed false positive CNV calls to the analysis.

In a similar re-analysis of a UK-only population sample we again saw similar results, albeit with a reduced significance attributable to reduced power. This reassured us that our results were not confounded by population stratification.

Taken all together these results are notable not so much for the differences between the cases and WTCCC2 controls, which although statistically significant are quite small, but particularly for the difference in deletion frequency between the cases and screened control cohort. Given that the cases and screened control samples were both derived from blood samples and genotyped on identical Illumina arrays at the same laboratory at the same time, we were confident that the difference in deletion burden between these two cohorts was real. In counter-balance however, it must be pointed out that the number of screened control samples was quite small (n=348), and this resulted in confidence intervals that were wider than if the sample was larger. Follow up analysis of other screened control samples would be warranted to confirm our findings in this group.

We also analysed regions of the genome that had previously been identified as harbouring CNVs associated with schizophrenia, using a list of loci previously collated by Kirov(Kirov, 2010). Since there is some evidence that different psychiatric diseases can be associated with deletions or duplications at the same locus(Shane E McCarthy et al., 2009), we analysed the frequency of both deletions and duplications in our sample. When deletions and duplications were

considered together, we found a significant difference between cases and screened controls (p=0.019), but not WTCCC2 controls. However no particular locus showed statistical significance on its own. Specifically (table 2.20) we saw 2 instances of the 1q21.1 deletion (0.07% of RDD cases vs. 0.2% of schizophrenia cases(Kirov, 2010)), 2 instances of the 2p16.3 deletion (0.07% of RDD cases vs. 0.2% of schizophrenia cases(Kirov, 2010)), 10 instances of the 15q11.2 deletion (0.37% of RDD cases vs. 0.6% in schizophrenia cases(Kirov, 2010)), 0 instances of the 15q13.3 deletion (0% of RDD cases vs. 0.2% schizophrenia cases(Kirov, 2010)), 7 instances of the 16p13.1 duplication (0.26% of RDD cases vs. 0.3-0.5% schizophrenia cases(Kirov, 2010)), 3 instances of the 16p11.2 duplication (0.04% of RDD cases vs. 0.3% of schizophrenia cases(Kirov, 2010)) and 3 instances of the 22q11.2 deletion (0.11% of RDD cases (frequency not stated in Kirov paper)). Note that in these cases we have counted CNVs affecting the consensus region and a genic area thought to mediate the aetiological effect (e.g. the neurexin 1 (NRXN1) gene at 2p16.3).

Therefore it seems from this evidence that some CNVs previously implicated in schizophrenia are also implicated in RDD, for example the 16p13.1 duplication, whilst others are not, for example the 15q13.3 deletion. However our numbers are too small for statistically significant association.

Subsequently we used PLINK(Purcell et al., 2007) to analyse our data, as this program conveniently analyses differences in not only the proportion of samples with a CNV of a particular type, but also the number of CNVs per sample, the average total CNV burden per sample and the average CNV size per sample.

When cases were compared to screened controls the number of CNVs per sample was significantly increased (1.58 vs. 0.92, $p<1.0x10^{-4}$), and this was driven by deletions (0.99 vs. 0.41, $p<1.0x10^{-4}$). Average total CNV burden per sample was also significantly different (563.5kb vs. 400.4 kb respectively, $p=2.0x10^{-4}$). This was driven by deletions (475.7kb vs. 275.7kb, $p<1.0x10^{-4}$) but not duplications (418.0kb vs. 370.8kb, p=0.11). When cases were compared to WTCCC2 controls the number of CNVs per sample was significantly increased (1.58 vs. 1.41, p=0.0028), again principally driven by deletions (0.99 vs. 0.84, p=0.0033) although the absolute difference is quite small. However for average total CNV burden per sample there was no significant difference overall (563.5kb vs. 558.8kb respectively, p=0.41), and no significant difference in total deletion CNV burden (475.7kb vs. 464.8kb, p=0.33) or total duplication CNV burden (418.0kb vs. 453.6kb, 2-sided p=0.42). It is possible in this instance that cell line derived CNVs not removed by QC procedures may be obscuring a true difference in some metrics.

We then went on to analyse singleton CNVs, which are defined as those that only occur once in a dataset, and thus by definition rare events. When we compared cases to screened controls the number of singleton CNVs per sample was significantly increased, driven by singleton deletion CNVs (0.20 vs. 0.13, p=0.017). The total singleton deletion CNV burden was also increased (251.8kb vs. 195.7kb, p=0.032). However the proportion of samples with a singleton deletion CNV was not significantly different (0.14 vs. 0.13, p=0.29). Thus cases are more likely to have more than 1 singleton deletion CNV than screened controls. When we compared cases to WTCCC2 controls the proportion of cases

to have a singleton event was significantly different (0.10 vs. 0.078, p=0.0010), as was the number of singleton CNVs per sample (0.13 vs. 0.096, p=0.0020), but not, in this case, the total CNV burden per sample (220.9kb vs. 226.4 kb, 2 sided p=0.72). Note that the numbers of samples with singleton CNVs reduces as total sample number increases, and fewer CNVs become defined as singleton across the dataset. Thus cases are more likely to have a singleton deletion CNV than WTCCC2 controls, but the average burden across samples is similar. Thus, overall, a case is more likely to have more than 1 singleton CNV of a defined size than a control, which is consistent with our previous observations.

Finally, we validated 40 calls within our sample set on a high density oligonucleotide CGH array. The 40 CNVs were comprised of 28 deletions (length: mean=329,433bp, range=106,370-1,405,099bp and number of markers: mean=94, range 23-344) and 12 duplications (length: mean=481,894bp, range=126,731-1,394,315 and number of markers: mean=115, range=19-343). All 40 CNVs were validated. This serves as some reassurance of the integrity of our calls, although it must be noted that the likelihood of confirmation is greater in calls that are larger and made with more markers and only one of these samples was from the bottom 10% of samples via QC metrics (we found that samples with poorer quality metrics were less likely to have DNA available for follow up). The results from this array also do not rule out the possibility of false positive calls in regions of the genome it did not cover.

7.2.2 Common CNVs

Chapter 3 concentrates on an analysis of common copy number variation in our sample. Generally speaking, common variation has been less intensively studied in association studies with disease, because their prevalence makes it logically less likely that they will confer a strong effect on morbid processes that affect fecundity. A study by Conrad et al. suggested that whilst some common CNV loci were associated with disease states, the heritability void left by GWA studies was not accounted for by common CNVs(Conrad et al., 2010). Craddock et al.(Craddock et al., 2010) have previously reported tagging of a variety of common copy number variants detected by array CGH with known SNPs. That is, the CNV segregates with the SNP, which in this context is called a 'tagSNP'. We decided to analyse common CNVs in our sample by using tagSNPs that were present within our consensus marker set.

The relative proportions of common CNVs varies amongst different populations (Lander & Schork, 1994). To account for the potential for population stratification we used only samples from the UK population and corrected for the first two principal components derived using Eigenstrat(Price et al., 2006). We then used 516 SNPs derived from our cleaned marker set which tagged ($r^2>0.8$) common CNVs and ran an association analysis under a logistic regression model using PLINK(Purcell et al., 2007). We found no results that survived Bonferroni correction for multiple testing ($p=9.69x10^{-5}$) when we compared cases to screened controls or WTCCC2 controls. We found that one tagSNP (rs12035407) was associated in each cohort with a p value of < 0.05, and tagged a gene 14.6kb

downstream from the SNP, FMN2 (formin 2). This gene has not been associated with neuropsychiatric disorders.

Our results are perhaps reassuring. No genome wide association study of SNPs has replicated findings in depression(Muglia et al., 2010; Rietschel et al., 2010; Shi et al., 2011; Wray et al., 2012), including one done using this sample(Lewis et al., 2010), but if common CNVs are tagged by SNPs then previous studies are likely to have highlighted them.

7.2.4 Chromosome 22q11.2

We next went on to investigate the 22q11.2 region more closely by running a small proportion (n=325) of our GWAS samples on a high density oligonucleotide comparative genomic hybridisation array (aCGH). aCGH microarray platforms have advantages over SNP microarrays because the reference source used in aCGH is DNA, rather than an inferred reference calculated from all samples. Furthermore, probe coverage is often much denser than SNP arrays. 22q11.2 is an interesting region as it is flanked by unique regions of segmental duplications and, of all regions of the human genome, is most frequently implicated in disorders involving genomic rearrangement, suggesting it is under current selection in the genome(Dunham et al., 1999; Shaikh et al., 2000). A 3MB deletion at 22q11.2 is the most commonly seen genomic deletion syndrome, probably occurring in about 1 in 3,000 live births(Burn & Goodship, 1996). Psychiatric disorders(Karayiorgou et al., 2010; Niklasson & Gillberg, 2010), including mood disorders(Jolin et al., 2009; Papolos

et al., 1996) are unusually highly prevalent in individuals with 22q11.2 deletions, making it a reasonable area for focussed study.

Initially we took our PennCNV calls, relaxed our CNV inclusion criteria, and re-analysed our rare CNV calls over the 22q11.2 area. We found that the proportion of samples harbouring a rare deletion was significantly increased in cases compared to controls ($p=0.0014$), however the average total rare deletion size was not significantly larger ($p=0.85$), although it was for rare duplications ($p=0.031$). The interpretation of this finding is that whilst a case is no more likely to have a rare duplication than a screened control, if they do the duplication is more likely to be large. This is driven by a number of large rare duplications in our cases over 22q11.2 which do not appear in screened controls, but due to low screened control sample numbers, this is not statistically significant in itself, and therefore the relevance of this finding is unknown. Large 22q11.2 duplications have been associated with disease states previously(La Rochebrochard et al., 2006; Ou et al., 2008; Portnoī, 2009; Yobb et al., 2005) although the phenotype is generally milder than deletions. We validated 207 calls made by PennCNV. Of these, 73% were confirmed as true positive and 26% were confirmed as false positive (1% could not be verified). Calls that were not validated were significantly more likely to be shorter ($z=-5.65$, $p>|z|<0.0001$) and made with less markers ($z=-3.77$, $p>|z|<0.0002$).

We then extracted copy number calls direct from the aCGH data by setting limits for the detection of a deletion and duplication based on metrics derived from all segments called by the program DNAcopy(Olshen et al., 2004), which is a

popular segmenting methodology for aCGH data. We again analysed our data with PLINK, stratifying our calls into common and rare variants, and analysing 4 metrics produced by PLINK. The wide variability seen in the absolute measures in the different metrics, but the failure to detect a significant effect, suggests that we did not have sufficient power to detect an effect in this analysis.

Finally, we used our array CGH data to delineate the breakpoints of the large 22q11.2 deletion CNV seen in one of our cases. This CNV conformed to the usual nature of 3MB deletions seen in this region(Shaikh et al., 2000).

7.2.5 iPattern and QuantiSNP Calling Methodologies

In our final analysis chapter, we ran our data using two alternative CNV calling methods, QuantiSNP and iPattern. We sought to reinforce the evidence for our hypothesis that rare deletion burden was associated with cases by calling the chip data with two different methodologies for CNV detection. QuantiSNP implements a hidden Markov model, and in this sense is similar to PennCNV, however QuantiSNP uses an objective Bayesian method to dynamically set prior parameters and sets a prior expectation of how frequently the model should deviate from the state of normal copy number in a different manner to PennCNV(Colella et al., 2007). This is a feature which may reduce the amount of type 1 errors, although in balance may also increase the frequency of type 2 errors. iPattern uses a different approach to detecting CNVs and takes as its input not the normalised relative and allelic intensity ratios (LRR and BAF) but the normalised raw fluorescence values, which Illumina denote as ‘X’ and ‘Y’. GenomeStudio uses X and Y values to generate LRR and BAF values (see 2.3.2.4)

using the GenCall method to cluster all samples to derive canonical reference values. iPattern, on the other hand, performs clustering of markers within batches of similar samples created based on a pre-calculated metric of the variance of the X and Y values, itself representative of the signal to noise ratio (Dalila Pinto, personal communication). Deriving batches of similarly performing samples from noise to signal ratio metrics, and deriving canonical reference values from these batches may be a more accurate methodology than creating a single reference value for each marker from all samples. However, the iPattern methodology remains unpublished in a peer-reviewed journal at the time of writing.

The rate of concordance between different CNV calling methods is low(Dellinger et al., 2010; Pinto et al., 2011; D. W. Tsuang et al., 2010; Winchester et al., 2009). Therefore it logically follows that CNVs called by more than one method are more likely to be true positive than those called with only one method. Therefore we also generated an intersected CNV call set based on a consensus of boundaries taken from overlapping calls made by QuantiSNP and iPattern in the same sample, with additional validation (presence/absence) by PennCNV. This methodology has been used in previous high profile CNV studies(Pinto et al., 2010). Whilst intersection of calls is likely to result in a lower false positive rate it may, in logical consequence, result in a higher false negative rate.

CNV call accuracy is also affected by sample quality. To generate a high QC sample list with samples at a lower probability of containing artefacts that might generate false calls, we further refined our sample QC metrics by taking into

account the number of calls made with iPattern and QuantiSNP and also taking into account metrics that identify samples with wavy signals and samples with excessive numbers of spurious LRR values. This resulted in a list of high quality samples that could be used for phenotypic association analyses.

To analyse our data from each method we used PLINK, as in our burden analysis in chapter 2. We modified our original analysis pathway used in chapter 2 to better account for calls falling within areas of copy number polymorphism. As such our calculations and results from the PennCNV method in this chapter differ slightly from those in chapter 2. Within our analysis in chapter 2 we realised that some regions of copy number polymorphism, particularly those with a proportion of occurrence across our dataset of between 1-3% being included in the analysis. This may introduce an element of bias by including areas of copy number polymorphism which vary according to population. In fact in our analysis restricted to a UK only sample set we did not see a significant deviation in results of our rare CNV analysis, so this issue may not be of importance. The issue of population stratification in CNV analysis will be discussed further below.

With the PennCNV method, as expected, we found an increased proportion of samples with rare deletion CNVs in our standard QC cohorts when cases were compared to screened controls ($p=0.01$, 40.3% vs. 33.6%) and WTCCC2 controls ($p=0.0004$, 40.6% vs. 36.6% respectively). Case frequencies of CNVs vary slightly (in this example 40.3% vs. 40.6%) because of the method used for removing regions of copy number polymorphism. Within our high QC analyses we did not find a significant difference in the frequency of cases with rare deletions when

compared to screened controls (p=0.18, 34.6% vs. 31.9%). Striking here is the reduction in frequency in samples with deletion CNVs in our case cohort between our standard QC and high QC analyses (40.6% vs. 34.6%). In the same analysis comparing the frequency of rare deletion CNVs between cases and WTCCC2 controls we did however find a significant difference between our cohorts (p=0.0066, 36.1% vs. 33.2%). Whilst the failure to find a significant difference in the proportion of high QC cases and screened controls with rare deletions called by PennCNV is surprising, other metrics did indicate a significant difference. The average rare CNV deletion frequency was significantly increased between high QC cases and screened controls (p=0.002, 0.59 vs. 0.39), the average total CNV deletion burden was significantly increased (p=0.0061, 361kb vs. 258kb) and the number of genes spanned by deletion CNVs was also significantly increased (p=0.024, 0.13 vs. 0.084).

With the iPattern method we generally found support for the associations made with data from the PennCNV method, although sometimes these failed to reach statistical significance. Significant associations were seen for the deletion CNV event rate per person (p=0.0091, 0.53 vs. 0.38) and the total CNV event distance spanned per subject (p=0.0008, 436kb vs. 266kb). iPattern data showed that cases were more likely to have a rare duplication event than screened controls (p=0.041, 0.57 vs. 0.45) in the standard QC cohort. iPattern data also generated similar findings when cases were compared to WTCCC2 controls, with the absolute differences between cohorts being less but significance values being higher owing to the larger sample size. Within the high QC analysis, iPattern data showed a significant association between the rare deletion CNV event rate when

cases were compared to screened controls (p=0.029, 0.47 vs. 0.36) and WTCCC2 controls (p=0.0001, 0.60 vs. 0.44) although notable in this comparison was the large difference between rates between cases processed with screened controls, and cases processed with WTCCC2 controls. This likely derives from the copy number polymorphism exclusion issue already described. Similarly iPattern showed that there was a significant difference between cases and screened controls in terms of the total CNV distance spanned per subject (p=0.0043, 408kb vs. 269kb), which was also seen when cases were compared to WTCCC2 controls (p=0.0001, 499kb vs. 364kb). Again the difference between cases when processed with the screened controls, and when processed with the WTCCC2 controls, was notable. iPattern did not support the finding with PennCNV data that rare deletion CNVs were more likely to involve more genes in high QC cases compared to WTCCC2 controls, although this narrowly failed to reach statistical significance (p=0.062, 0.11 vs. 0.094). The number of CNVs involving at least one gene was however significantly associated with iPattern data in the same cohorts (p=0.036, 0.097 vs. 0.084).

Throughout both our standard QC and high QC analyses the QuantiSNP method demonstrated no significant differences between either cases and screened controls or cases and WTCCC2 controls that supported results from PennCNV or iPattern data. Notably, data from the QuantiSNP methodology in our standard QC cohorts suggested that there were more rare deletion events in WTCCC2 controls than cases (p=0.015, 0.33 vs. 0.37) and that similarly the proportion of samples with a rare deletion CNV was also increased (p=0.012, 0.27 vs. 0.30). The high QC analyses data from QuantiSNP generated no significant differences between

cohorts in any metric. This was a surprising result. We noted that QuantiSNP generally called fewer CNVs than either PennCNV or iPattern (rate of all rare CNVs called in high QC analyses = 1.03, 0.92 and 0.76 CNVs per sample for PennCNV, iPattern and QuantiSNP methods respectively), but this does not necessarily explain this observation.

Within our intersected call set we repeated our analyses above (shown in the appendix). Given that the QuantiSNP method had produced no significant results, and that our intersected call set required calls to be made by QuantiSNP in order to be included in our analysis, we did not expect to find significant results using this dataset. No analysis showed any significant difference between cases and screened controls or WTCCC2 controls.

We attempted to cast some light on the natural question to arise from our results; are we observing a true difference in CNV burden between cases and controls, or are our results confounded by false positive calls that tend to be made by some methods but not others? We validated calls made by each method separately in the 22q11.2 region in those samples that were run on the CGH array described in chapter 4. If PennCNV was contributing false positive deletion calls, we might expect to see an increased rate of false positive deletion CNVs from the PennCNV call set. In fact our validation analysis revealed only 1 (1.4%) false positive deletion CNV called by PennCNV and 2 (2.8%) by QuantiSNP. iPattern fared worse in this respect (9 false positive deletion CNVs (12.0%)).

Finally we decided to reanalyse the regions of the genome previously implicated in schizophrenia, which we had previously looked at in 2.4.5, using our high QC

intersected call set. We found that some large CNVs, especially duplications, had been removed from this dataset. On further investigation this was not necessarily due to the samples being removed through our high QC thresholds, but rather these CNVs had never been called by the iPattern and QuantiSNP methods. This may mean that the intersected call set is over-conservative and underlines the delicate balance to be struck between type 1 and type 2 errors in CNV calling methodologies.

7.2.6 Phenotype Analyses, Sex Chromosome Syndromes and Specific CNVs

In our final analysis chapter we performed association analyses of CNV burden against phenotype data derived from all case studies (GENDEP, DeNT and DeCC) included in the GWA study. We hypothesised that CNV deletion burden may be associated with lower age of onset of disorder, as others have provided evidence for in bipolar disorder(Malhotra et al., 2011; Priebe et al., 2011), and that there would be a relationship between CNV burden and symptom pattern and/or personality traits. We extracted data on age of first onset of depressive disorder, the duration of worst episode, SCAN items which we used to reconstruct 3 putatively familial factors previously described by Korszun et al(Korszun et al., 2004), and trait psychoticism, neuroticism and extraversion scores calculated from data taken from the Eysenck personality questionnaire(H. Eysenck & Eysenck, 1964).

We did not find any significant associations between any of our phenotypic measures and any measure of rare CNV burden or type. The most notable observation was a trend ($p<0.05$) for trait neuroticism scores to be associated

with rare deletion burden that ran across data derived from the PennCNV method, the iPattern method and the intersected call set, but not the QuantiSNP method.

We were surprised to find no associations in this dataset that survived Bonferroni correction for multiple testing. This may mean that CNV burden is related to more general depression symptoms not captured by our sub-phenotypes, or that CNV burden is related to another phenotype that we do not have data for, but is related to our case sample, for example cognitive ability.

We next analysed the frequency of sex chromosome abnormalities in our dataset by visually analysing plots of the X and Y chromosome in case samples where the phenotypic gender was different to that inferred from the hetrozygosity of the B allele frequency calculated by PennCNV. We showed that the frequency of Klinefelter's and Turner's syndrome in our dataset was not significantly different to the frequency observed in a large series of sequential liveborn neonates. Two out of three cases of Turner's syndrome in our cases were 45,X/46,XX mosaics. We also showed evidence of a sample with a diploid/triploid mosaicism, although in a study of recurrent depression this is much more likely to be a somatic event than a congenital one, as cases of congenital diploid/triploid mosaicism usually have multiple physical anomalies and learning difficulties(van de Laar et al., 2002).

In the final section of this chapter we presented a series of case studies of CNVs that had been selected based on their size, position or gene coverage. Some cases

with interesting CNVs had also been followed up with psychometric testing and neuroimaging.

In particular we were fortunate that the case with the large 22q11.2 deletion associated with cardiac deficits, facial dysmorphism and psychiatric disturbance had been followed up in this manner. Whilst he demonstrated a long history of social anxiety and depression, he displayed a normal IQ of 101 and an MRI scan which showed no major cranio-facial or CNS pathology.

A second case had a rare duplication affecting an isoform of *DISC1* (disrupted in schizophrenia 1). This was named after the discovery of a large Scottish pedigree with a rare translocation with a breakpoint affecting it(St Clair et al., 1990). In fact the translocation did not segregate solely with schizophrenia, with a high prevalence of psychotic, affective and anxiety disorders all noted within this pedigree. This case had little clinical history available, although she was known to have suffered at least three severe depressive episodes with one requiring hospitalisation. She remained chronically depressed at follow up. She also demonstrated a normal IQ but with a significant discrepancy between her verbal and performance scores. The CNV covered exons 10 and 11 of the gene, and did not cover the translocation breakpoint. *DISC1* has been associated with schizophrenia, bipolar disorder and schizoaffective disorder(C. A. Hodgkinson et al., 2004) but not depressive disorder. Interestingly, determination of regional expression of *DISC1* by in-situ hybridization in the primate brain localizes it to the dentate gyrus of the hippocampus, which may have a functional role in the aetiology of depressive disorder(Sahay, Drew, & Hen, 2007).

A third case had a rare deletion in the gene *GPC5 (*glypican 5). This deletion was also seen in the population controls from the WTCCC2 (n=3), however this gene is interesting because it is highly expressed in the CNS and also has variants linked to multiple sclerosis(Baranzini et al., 2009), a demyelinating disorder that results in a high prevalence of comorbid depressive disorder (37-54%)(Patten, Beck, Williams, Barbui, & Metz, 2003). This case had a particularly severe clinical course, with an early age of onset in the teenage years, a chronic course refractory to medical and psychological treatment, multiple suicide attempts and two courses of electro-convulsive therapy, which is usually reserved as the last clinical line of treatment for particularly intractable cases of depressive disorder. Of note, however, was this case's relatively high IQ although with a notable discrepancy between her verbal IQ of 129, and performance IQ of 108.

A fourth case harboured a particularly large deletion CNV, over 9MB in size. The proximal breakpoint interrupted the first exon of *PCDH9* (protocadherin 9). This gene is ubiquitously expressed in both developing and adult brain, and the gene encodes a cadherin-related neuronal receptor that localises to synaptic junctions(Pruitt, 2004). The region this CNV occupies is gene-sparse, and of the 9 genes affected by it, the functions were either entirely unknown, or the transcript did not localise to brain regions. This case had suffered from recurrent depressive episodes and also suffered from asthma, but otherwise had no other remarkable medical history. On entry to the study she scored 42 on the BDI (severe depression) and was treated with Nortriptyline, a tricyclic antidepressant. At follow up 12 weeks later she scored 18 on the BDI (mild-moderate depression). No other clinical details were available.

A fifth case harboured a duplication CNV interrupting exon 11 of *CACNA1C* (calcium channel, voltage-dependent, L type). This gene encodes an alpha-1 subunit of a voltage-dependent calcium channel. This case had an early age of onset (14 years) and a high trait neuroticism score (22). A CNV in CACNA1C is interesting, as a meta-analysis(Ripke et al., 2011) has implicated a SNP in this gene with bipolar disorder, twin studies suggest that bipolar disorder and depression share a substantial heritable component(McGuffin et al., 2003) and a further GWA study associated the previously implicated SNP with risk of schizophrenia and recurrent depressive disorder(Green et al., 2009).

A sixth case harboured a CNV on chromosome 2 which deleted one copy of *SLC5A7 (*solute carrier family 5 (choline transporter)), the protein of which enables the uptake of choline into cholinergic neurones for the synthesis of acetylcholine. *SLC5A7* is moderately highly expressed in the cingulate cortex and prefrontal cortex. The anterior cingulate cortex has been particularly implicated in the neurobiology of emotion and affective processing(Decety & Jackson, 2004) and may have a functional role in depression. A SNP in this gene was linked to a subphenotype of depressive disorder in a Canadian cohort(Hahn et al., 2008) and the role of acetylcholine in mood disorders has been of interest since the observation that drugs that block the enzyme cholinesterase, which breaks down acetylcholine, can induce depressive symptoms(Janowsky, El-Yousef, & Davis, 1974).

A seventh case with an early age of onset of depressive disorder (16 years) harboured a large singleton duplication CNV over chromosome 10, which

interrupts the first exon of the *NRG3 (*neuregulin 3) gene, and duplicates the *PCDH21* (protocadherin 21) and the GRID1 (glutamate receptor, ionotropic, delta 1) genes. *NRG3* is a particularly interesting gene as it is particularly expressed in the prefrontal cortex and amygdala. It has also been implicated in other psychiatric disorders in a variety of studies, with linkage to the region (10q22-3) first demonstrated in schizophrenia families(Fallin et al., 2003; S. V. Faraone et al., 2006) and further evidence from recurrent deletions in the area segregating in individuals with cognitive and behavioural problems including autism and hyperactivity(Balciuniene et al., 2007). More recently a candidate gene analysis implicated *NRG3* (neuregulin 3) in a large study of schizophrenia probands(Y. Wang et al., 2008) and finally a CNV study implicated a duplication CNV affecting *NRG3* (neuregulin 3) in one family(Xu et al., 2009). Unfortunately we had no more clinical information on this case.

An eighth case from a woman with the highest trait neuroticism score in our sample (23) and an early onset of recurrent and severe depressive episodes harboured a CNV deleting, amongst other genes, one copy of *GRIA4* (glutamate receptor, ionotrophic, AMPA 4 isoform), the *GUCY1A2 (*guanylate cyclase 1, soluble, alpha 2) gene and interrupting the *PDGFD* (platelet derived growth factor D isoform 2) gene. The *GRIA4* gene is particularly interesting as it has a previously reported association with schizophrenia(Makino et al., 2003), although other studies have been negative(Guo et al., 2004). The *CASP (*caspase) genes also reside over this region, although there is little current evidence to suggest their role in psychiatric disorders.

A ninth case, again with a history of recurrent and severe depressive episodes without psychotic symptoms since her early teenage years, was found to have a singleton deletion CNV on chromosome 12 deleting, amongst others, the genes *SLC6A15 (*solute carrier family 6, member 15 isoform 1), *NTN (*neurotensin/neuromedin N preprotein) and interrupting MGAT4C (UDP-N-acetylglucosamine: alpha-1,3-D-mannoside). The *SLC6A15* gene is particularly interesting in a case of depression, since a recent study in the journal Neuron associated a SNP in this gene with a cohort of cases with recurrent, severe depressive disorder, and the risk allele (AA) with a down-regulation of gene expression in the hippocampus and associated reduction in size(Kohli et al., 2011). From the GWAS SNP data, this individual was genotyped as GG. The hippocampus is a region known to be important in the aetiology of depressive illness(Frodl et al., 2002). This CNV also deletes one copy of the gene *NTN (*neurotensin/neuromedin N preprotein), an interesting neurotransmitter/neuromodulator that is widely distributed throughout the central nervous system, linked intimately to dopaminergic neurotransmission and may be involved in the aetiology of schizophrenia and antipsychotic response(Binder, Kinkead, Owens, & Nemeroff, 2001). This gene is highly expressed in the amygdala, hypothalamus and thalamus.

Finally four cases with deletions and duplications of a region on chromosome 15q13 previously associated with autism(Babatz, Kumar, Sudi, Dobyns, & Christian, 2009; Guilmatre et al., 2009) and schizophrenia(Kirov et al., 2007; Need et al., 2009) were shown. This region is interesting because it carries the *APBA2* (amyloid beta A4 precursor protein-binding) gene, which is highly

expressed in all areas of the brain, including the prefrontal cortex, cingulate cortex and amygdala. Also called Mint2, the protein product of this gene is a neuronal adaptor protein that binds directly to neurexins as part of a multi-protein complex that may act as a facilitator of neurotransmitter release from synaptic vesicles(Biederer & Südhof, 2000; Dulubova et al., 2007). However this gene has not been implicated in affective disorders before.

7.3 Methodological Discussion

In this section we will discuss the advantages and limitations of our analysis, with a view to achieving a balanced discussion of this project.

7.3.1 Samples

A major advantage of our study was the large number of case samples. The study cohorts contributing to our case samples in this analysis combine to make one of the largest white European depression cohorts globally. However since the samples were collected at many different sites, sample heterogeneity may be a problem. Whilst these samples were interviewed with SCAN(Wing et al., 1990), which is recognised as a standard and exhaustive neuropsychiatric assessment, the diagnosis of depression is unlikely to represent a single underlying neurobiological problem. To attempt to narrow the phenotype, individuals were only selected for study if they had suffered recurrent depression of at least moderate severity. Similarly, those with depressive disorder secondary to drug or alcohol misuse and with histories of mood incongruent psychotic symptoms were excluded. However narrowing the phenotype need not lead to a narrowing

of genotype. Lewis et al.(Lewis et al., 2010) have published a GWA study of SNP variants in a subset of individuals from this cohort, finding no replicated association results that survive correction for multiple testing. Indeed no GWAS in depression has yet achieved this(Muglia et al., 2010; Rietschel et al., 2010; Shi et al., 2011; Sullivan et al., 2009; Wray et al., 2012). Within our cohort the age range was wide (18-89). Individuals with an earlier onset of affective disorders (≤18 years) may have a higher deletion CNV burden than those with a later onset (>18 years)(Malhotra et al., 2011; Priebe et al., 2011) although in our own regression analyses we did not find evidence to support this in our dataset, perhaps because we considered age of onset as a quantitative, rather than a dichotomous, trait. Nonetheless, the point remains that heterogeneity of samples may obscure findings in subgroups.

The WTCCC2 controls were a large population control set, which in practice means that those who suffer from psychiatric illnesses such as recurrent depressive disorder, bipolar disorder or schizophrenia, need not be excluded. The advantage of this group is the large sample number (n=5,619), however half of these samples (those derived from the 1958 birth cohort) were derived from lymphoblastoid cell lines. The process of transformation is known to introduce genetic variation(Redon et al., 2006) although samples with large aneuploidies will have been excluded by our QC methods. Nonetheless the introduction of cell line derived copy number variants may artificially increase the burden of CNVs in the WTCCC2 controls and obscure true differences in burden.

A disadvantage of our study was the low number of screened control samples. When analysing the screened control samples from the DeCC cohort, which were derived from cheek swab DNA, we found unacceptably high numbers of artefactual calls which ruled out a meaningful genome wide analysis. We did go on to use these samples in analyses on focussed regions of the genome with large CNVs previously associated with schizophrenia in chapter 5, however in this case it was possible to visually validate every call. The screened control samples we used in our burden analysis were derived entirely from the bipolar association case control study(Gaysina et al., 2009). It should be noted that control samples from this study were derived in the main from staff and students from King's College London and by media advertisement. This may represent a particularly high functioning group of individuals, although we did not have measures of IQ or cognitive ability in this cohort.

A further disadvantage in our study was the non-availability of DNA samples from parents of probands. We have illustrated an interesting case series in this thesis, but without DNA samples from biological parents it is impossible to know if these CNVs were inherited or de novo germ line events. De novo events are interesting, especially if the biological parents of a proband with disease do not exhibit the disease themselves. Several groups have presented evidence that sporadic cases of apparently heritable diseases may in part be explained by de novo events (Malhotra et al., 2011; Sebat et al., 2007; Xu et al., 2008). Similarly if a CNV in a proband is seen in an affected parent but not in an unaffected relative, then the case for an association is reinforced. Future collections of samples looking for the role of rare variants in disease states would be strengthened by

the collection of specimens from the biological parents of probands, although this can be a time-consuming exercise.

Whilst all our samples were of exclusively white European parentage this does not necessarily rule out problems with population stratification. The GWAS of depression performed with this cohort restricted its analysis to a UK only sample subset to account for this problem(Lewis et al., 2010). It is not thought that rare CNV analysis is adversely affected by population stratification, if only because rare CNVs are, by definition, more likely to be deleterious and therefore tend to die out within populations within an evolutionarily short space of time, although a recent simulation study suggested that populations could potentially display spatial structure with respect to rare variants(Mathieson & McVean, 2012). In our rare CNV analysis in chapter 2 we additionally analysed a cohort of individuals with purely UK ethnicity and observed a similar result. We thought it was therefore reasonable to conclude that our results were not overtly biased by population stratification, although an element of this cannot be ruled out. In particular a part of the analysis pipeline in chapter 2 can be criticised for failing to exclude some regions of copy number polymorphism where calls do not precisely overlap with each other. In chapter 5 we took a more robust approach to excluding areas of polymorphism, and whilst our association statistics were somewhat different, the overall association appeared to be robust. In our common CNV analysis presented in chapter 3 we used a UK only sample set which had also been analysed with Eigenstrat to produce principal components to correct for population stratification. We found no significant associations

using this data, probably suggesting that our UK samples were relatively homogeneous after correction with the major principal components.

7.3.2 Genotyping

Our samples were sent to the Centre Nationale De Genotypage (CNG) for genotyping on the Illumina 610 Quad array. The laboratory is a purpose-built genotyping facility with full laboratory information management systems (LIMS) control and is very experienced in processing Illumina data. Notwithstanding this however the laboratory did not release the Illumina idat files to us, but rather insisted on supplying final reports generated by Illumina's GenomeStudio application. We saw this as a problem, and made strenuous efforts to obtain the idat files to process within GenomeStudio ourselves. This was because we wanted to be sure that both our case/screened control samples and the WTCCC2 samples were processed using GenomeStudio in an identical manner. We were however repeatedly denied access and eventually had to process the final reports with reassurances that they had been clustered and processed according to Illumina's standard guidelines, which is how we processed the WTCCC2 controls. However we also learned subsequent to our analysis that the CNG do not cluster samples based on a reference file generated from the samples themselves, but use a predefined reference file derived from groups of samples genotyped at their own lab. It is impossible to tell whether or not this will have affected our CNV calling.

The Wellcome Trust controls were genotyped on a variant of the Illumina 1M beadarray. The samples were genotyped at the Wellcome Trust's Sanger centre

in Cambridge. The beadarray technology and Infinium HD assay used in the 610 Quad chip, on which our cases and screened controls were genotyped, and the 1M chip, which the Wellcome Trust controls were genotyped on, is identical except that the number of probes on the array differs and the variant 1M chip used by the WTCCC2 does not carry the monomorphic copy number probes seen in the 610 Quad chip. To account for this we created a consensus marker set in common between the two arrays. It is reasonable to criticise our decision to use a consensus marker set. Certainly if we had had access to a larger control group genotyped on the 610 Quad it would have been preferable to use this, but the majority of our screened control samples were derived from unsupervised cheek swabs, which turned out to have high numbers of artefacts that precluded their use in genome wide burden analyses. The addition of the WTCCC2 samples provided a large independent control set of samples derived from the UK with which to compare our cases too, but it is possible that the different density of probes on the 1M chip, when compared to the 610 Quad chip, may have affected our results, as well as the different time and location of genotyping.

We were however reassured in our initial analyses in chapter 2 that the likelihood of systematic bias between the three cohorts was relatively low, since the WTCCC2 controls, which were genotyped at a different laboratory to the cases and screened controls, had a CNV burden that appeared to fall in-between the cases and screened controls. If the genotyping conditions were systematically affecting the results we would have expected the screened controls and cases to both have had higher or lower numbers of calls relative to the WTCCC2, instead of a figure flanking them. In our subsequent analysis with

the QuantiSNP method in chapter 5 there was, in any event, no significant difference between the datasets.

7.3.3 CNV Calling Methods

We initially chose the PennCNV method because it was well-established, designed to be used with Illumina data, and easy to use. A recent paper by Pinto et al(Pinto et al., 2011) studied different arrays and calling methods for CNVs. They concluded that "Different analytic tools applied to the same raw data typically yield CNV calls with <50% concordance". This was an overall figure, and the authors also pointed out that reproducibility is generally better with Illumina arrays than Affymetrix arrays. Moreover, this observation was made on the basis of calling CNVs with a minimum contiguous marker threshold of 5 and a minimum size of 1kb. Our own minimum thresholds were 10 markers and 100kb respectively, so this analysis may not necessarily reflect our own, nonetheless the reproducibility between methods is a concern, and this is sharply reflected by the differing results we achieved with PennCNV, iPattern and QuantiSNP. In Pinto et al.'s paper they analysed the PennCNV, iPattern and QuantiSNP methods with respect to Illumina datasets. The QuantiSNP method consistently called less CNVs when compared to PennCNV and iPattern, which is consistent with our own data. Although they concluded that Illumina data generally performs better than other platforms, the between-replicate reproducibility and between-method reproducibility is notably poor. Given this, it is perhaps unsurprising that our different call sets do not produce entirely similar results. Many published studies including our own have relied on calls made with just one method(Grozeva et al., 2010; McQuillin et al., 2011; Rucker et al., 2011; D. Zhang

et al., 2009a). It is possible that re-analysis of this data with different methods would lead to different results, complicating interpretation. Whilst some have called for a consensus on CNV calling methodologies(Scherer et al., 2007), the reality of rapidly changing array platforms and tools for analysis makes this an unlikely reality.

Both the PennCNV and QuantiSNP methods use quite similar methods for detection, yet both methods have produced quite different results in our analysis. The results from the iPattern method lie in-between. As we discussed in chapter 2, hidden Markov models (HMMs) calculate the probability of copy number states being present given the data and the model's set prior parameters. Given that the data fed to the model is similar, perhaps the prior parameters affect CNV calling.

Amongst the prior parameters set within the HMMs is a prior expectation of the degree of conformity of the model to the null state, which in this instance is normal copy number. Put another way, this value dictates to what extent the model expects to find copy number variation. QuantiSNP sets a parameter called "longChromosome", which is set at $2x10^6$ bp. In other words the model expects normal copy number to be present, on average, for 2MB at a time. PennCNV treats each marker position as being equally likely to be within a CNV region, and sets a background probability of variation from this assumption of 0.0001. This may be an important difference between the two methods that may influence the probability of false positive and false negative calls in different samples. The

iPattern method has not been subject to peer-review at the time of writing, and information on its prior parameters has not been made available.

The PennCNV and QuantiSNP methods also rely on pre-existing LRR thresholds, derived from existing data, for calling CNVs. Within QuantiSNP this is derived from data on experimentally validated regions of known copy number variation from Illumina Human-1 and HumanHap 300 genotyping arrays(Colella et al., 2007). These arrays are based on an earlier Illumina technology than the Infinium HD arrays used in our experiments, and the default thresholds may be overly conservative. PennCNV, on the other hand, sets thresholds derived from Illumina Infinium genotyping chips similar to our own, which in turn may lead to more realistic results with our data. iPattern, which has not been subject to peer review, sets thresholds dynamically within batches of samples deemed to be similarly performing based on the variance of the X and Y values. It is possible that the results from the three methodologies used in this thesis are broadly reflective of the initial parameters used for setting and varying LRR thresholds for CNV calls. A future re-analysis of this work could include setting different parameters within QuantiSNP and PennCNV based on thresholds derived from samples in our cohort with known, large CNVs of different types.

In a another study of CNV calling methods, Dellinger et al.(Dellinger et al., 2010) studied the performance of seven methods for CNV detection, including QuantiSNP and PennCNV. The other methods compared in this paper are designed for use with data derived from Affymetrix chips and array CGH data. They found that the QuantiSNP method performed best with their data, with

PennCNV performing moderately well, but this was concluded largely from simulation studies. They found that the PennCNV method called the least CNVs, which is opposed to our data, whilst QuantiSNP called the most CNVs, however they set different thresholds to our own. The iPattern method was not analysed in this study.

Generally speaking, CNV calling methods are one of many 'black boxes' within the analysis process from sample to CNV call. Not enough information is provided within the original papers for the average reader to make a judgement about the relative advantages of each method and subsequent to publication numerous updates are usually made. More particularly, CNV calling methods tend to be designed to be used with particular arrays, if not particular manufacturers. It becomes a particularly tricky exercise to try and unpick the relative rationales for CNV call sets made with different methods in this situation.

Intersecting calls from different methods, and only taking calls made with multiple methods is logically likely to reduce the rate of false positive calls. However true calls may also be dropped if they are not made with all methods contributing to the intersection. It is impossible to quantify this without visually QCing every call that does not replicate between methods.

7.3.4 Sample and CNV Call Quality Control

Sample and CNV call quality control are critical in CNV analysis. The unfortunate reality of the current situation is that the thresholds for excluding both samples and CNVs vary according to sample origin, array manufacturer, array type, the type of CNV being called, the size of CNV being called and how many markers a

CNV call is based on. Methodologies vary between papers according to the nature of the datasets being processed, making methodological mimicry mostly meaningless and usually calling for a bespoke analysis pipeline for each dataset. This in turn makes replication between groups harder to achieve and the probability of replicated results lower. The efforts by the psychiatric GWAS consortium(Cichon et al., 2009) is likely to mitigate some of these problems, but ultimately it becomes very difficult to do meaningful analyses of CNV burden with samples of different origins run on arrays with different underlying technologies, with different probe densities and compositions and from different manufacturers. However large CNVs, in practice those greater than about 500kb, can usually be reliably detected in the majority of samples, and many analyses have reasonably concentrated on these CNVs.

Within our own samples we were fortunate to be able to restrict our analysis to those samples derived from venous blood. DNA from the 1958 birth cohort was derived from cell lines, and it is possible that artefacts introduced during cell line creation, which can include large aneuploidies, affected our results. In fact, we only found 7 samples with large aneuploidies, all from the 1958 birth cohort which we excluded as described in 5.3.1.4.

7.3.5 Statistical Analysis

Our initial analyses in chapter 2 could be criticised for reducing our call set to a binary 'sample-with-CNV vs. sample-without-CNV' dataset. Such an analysis may, for example, be confounded by poor quality samples contributing false positive calls in samples that otherwise may not have had a true CNV event. Such a

criticism is more an indictment of QC procedure than statistical methodology per se. The advantage of this approach is that it allows the calculation of odds ratios with confidence intervals that allow a broad idea of relative risk of disorder to be given. In the context of gene-disease association analyses this may be criticised as propagating the simplistic view that genes are directly associated with complexes diseases such as depression. Nonetheless the simplicity of the method is attractive and the results easy to disseminate.

PLINK(Purcell et al., 2007) implements max (T) permutation for CNV burden analyses, which allow the calculation of empirical p values for 1 and 2 sided tests of associations independent of an assumption that the model fits a defined distribution. Whilst the software is fast, easy to use, and calculates an abundance of useful metrics of relative and absolute CNV burden, it only reports mean values and empirical p values, rather than associated test statistics that may allow a more accurate assessment of the degree of effect (odds ratios, z scores, beta values etc.). We saw this particularly when comparing and contrasting the results from comparing our cases to our screened controls and WTCCC2 controls. Empirical p values for the comparison between cases and WTCCC2 controls were often large than those with screened controls, despite a much smaller actual difference.

7.3.5 A Final Summary and Discussion of Our Results

The genome wide analysis of CNVs with SNP microarray analysis may vary for numerous reasons, only one of which may be a true difference in burden between cohorts. We have discussed the problems inherent with samples,

genotyping and calling CNVs in previous sections, all of which may have had an effect on our results. At a more fundamental level, the problems with CNV calling from SNP microarrays may be summarised by considering the fact that the technology was never in fact designed with this in mind. SNP microarrays seek to make AA, AB and BB genotype calls from SNP markers based on the assumption of biallelic genotype. Thus the fluorescence data derived from each probe needs only to be binned into three possible states. Copy number calling relies on the binning of the same data into, usually, six separate states, which represent the range of possible copy number states from full deletion to a double copy duplication. CNV calling from microarray data is then essentially, an attempt to extract more information out of the same data, with all the problems that this entails. We were probably fortunate in this project to be using Illumina data, which is probably less susceptible to noise than other arrays, but nonetheless the problem of random noise affecting the attempt to derive more discrete states from the same fluorescence data remains.

We initially presented data that strongly indicated a stepwise difference in rare deletion burden between our cohorts. When we used three different methodologies for CNV detection this difference became less clear, with the QuantiSNP method failing to validate data from PennCNV and iPattern, possibly because the model's prior parameters were not optimised for our data. It may be that the PennCNV method generates an unacceptable number of false positive calls, or that the QuantiSNP method generates an unacceptable number of false negative calls. Our validation of calls with array CGH did not clarify this situation, because the rate of validation between PennCNV and QuantiSNP was very

similar. However our array CGH data only covers a small proportion of the genome, and to systematically validate CNVs from the different methods we would need to use whole genome array CGH.

It is the author's opinion, taking all the data presented in this thesis into account, and the experience of working with the data over the past three years, that the screened control cohort derived from the BaCC sample does have a significantly reduced deletion burden in comparison to the case cohort. The differences between the WTCCC2 cohort and the screened control and case cohorts were much less clear for the reasons discussed above, although often the large sample numbers resulted in statistically significant findings.

7.4 Our Findings in Relation to Other Studies in Affective Disorder

7.4.1 Depressive Disorder

Shortly prior to the end of this project Degenhardt et al.(Degenhardt et al., 2012) published a paper linking copy number variants at 16p11.2, 7p21.3, 15q26.3 and 18p11.32 with depressive disorder. 604 patients with depression and 1,643 controls from a variety of sources were genotyped on three closely related Illumina arrays (HumanHap 550, 610 Quad and 660W). Similar to our own methods they used a consensus marker set between the arrays and used the PennCNV and QuantiSNP methods for calling CNVs, analysing their CNVs with PLINK.

Contrary to our own findings, they did not find a change in burden between cases and controls. They set a more stringent threshold for inclusion of CNVs, which were kept only if made with ≥30 consecutive markers, a log Bayes factor (QuantiSNP) of ≥30 or a confidence score (PennCNV) of ≥30. This may explain the differences in our burden analyses.

Within our own results for regions 16p11.2, 7p21.3, 15q26.3 and 18p11.32, we do not see an excess of CNVs in our cases compared to either our screened controls or WTCCC2 controls. Within the 16p11.2 region we observe 2 cases with a deletion and 3 cases with a duplication, no CNVs in the screened control cohort and of the WTCCC2 sample we saw 3 samples with deletions, 1 sample with a full length duplication and 1 sample with a duplication spanning half the length of the normally seen CNV (derived from our PennCNV results), but these findings were not significant. Within the remaining regions we see a complex range of microdeletions and duplications, making comparisons between studies difficult.

Glessner et al.(Glessner et al., 2010) found an association between a duplication CNV at the SLIT1 locus and depressive disorder in 1,693 cases of depressive disorder when compared to 4,506 controls genotyped on the Perlegen 600k platform. They found 5 duplications in SLIT1 in their cases, but none in controls. They did not report a genome wide burden analysis, possibly because the Perlegen platform has a higher than average signal to noise ratio than other genotyping platform, making CNV calling more difficult. In comparison, we see 1

duplication of the SLIT1 locus in our cases, and none in our screened controls or the WTCCC2 controls (p=0.36, Fisher's exact).

7.4.2 Bipolar Disorder

Preliminary evidence of association of CNV loci at both 16p11.2(Shane E McCarthy et al., 2009) and 3q29(Clayton-Smith, Giblin, Smith, Dunn, & Willatt, 2010; Mulle et al., 2010; Quintero-Rivera, Sharifi-Hannauer, & Martinez-Agosto, 2010) have been published in bipolar disorder. Three studies have shown an enrichment of rare CNVs in bipolar disorder, especially in those with a younger age of onset(Malhotra et al., 2011; Priebe et al., 2011; D. Zhang et al., 2009a), however two other studies have not supported this finding(Grozeva et al., 2010; McQuillin et al., 2011).

We did not find an association between age of onset and deletion CNV burden in our analysis, although we could repeat this analysis by stratifying our data into categorical groups, as other teams have done, rather than performing the analysis with the quantitative trait. Age of onset in affective disorders, and probably psychiatric disorders in general, is arguably contentious. It is difficult to precisely define the point at which a pre-existing trait becomes a disease, and this may not be consistent between assessors. In this sense, analysis via a categorical model may be more meaningful.

Zhang et al. noted an increase in the proportion of cases with singleton deletions throughout the genome when compared to controls in their analysis of bipolar cases. Our PennCNV findings partially support this data. We found that the proportion of samples with a singleton deletion CNV was significantly higher in

cases than WTCCC2 controls (p=0.0010, 0.10 vs. 0.078, respectively), but that this did not replicate when cases were compared to screened controls (p=0.29, 0.14 vs. 0.13, respectively). However when the total burden of singleton deletion CNVs per sample was analysed, cases were more likely to have a higher burden than screened controls (p=0.032, 251.8kb vs. 195.7kb). Zhang et al. also noted CNVs in the GRM7 and LARGE genes, which had also been noted in Walsh et al.'s report in schizophrenia(Walsh et al., 2008). We observed 1 case and 1 WTCCC2 control, but no screened controls, with deletion CNVs interrupting GRM7, and 2 cases and 2 WTCCC2 controls, but no screened controls, with deletion CNVs interrupting LARGE.

We do not see any instances of the large deletion or duplication at 3q29 in any of our cases, screened controls or WTCCC2 controls (PennCNV data). This may indicate that this CNV is more specific to bipolar disorder and schizophrenia(Levinson et al., 2011), since our cases were screened for, and excluded if they had, manic or mood-incongruent psychotic symptoms and any history of mania or schizophrenia. Whilst we observed instances of the 16p11.2 CNV in our cases, as well as the WTCCC2 controls, none exhibited symptoms or history of mania or psychosis, including mood congruent delusions or hallucinations. However, again this is consistent with evidence that suggests that this particular CNV is associated with numerous psychiatric disorders(Degenhardt et al., 2012; Fernandez et al., 2010; Shane E McCarthy et al., 2009), as well as morbid obesity(Bochukova et al., 2009).

Priebe et al.(Priebe et al., 2011), in a study of 882 cases of bipolar disorder compared to 872 population controls, noted that 2 relatively common CNVs were significantly over-represented in cases of bipolar disorder with an age of onset ≤21 years, a 160kb microduplication on 10q11 and a 248kb microduplication on 6q27. We observed that the 10q11 polymorphism occurred in 34 out of 2,723 cases (1.25%), 5 out of 348 screened controls (1.44%) and 58 out of 4,828 WTCCC2 controls (1.20%) (PennCNV data). We observed the 6q27 polymorphism in 99 out of 2,723 cases (3.63%), 10 out of 348 screened controls (2.88%) and 167 out of 4,828 WTCCC2 controls (3.45%) (PennCNV data). No comparison yielded statistically significant results, although it is interesting to note that our percentage frequencies were reduced from those cited in Priebe et al.'s paper, probably because of differences between our sample cohorts.

7.5 Our Results in the Broader Field of CNVs in Psychiatric Genetics

Rare copy number variants (CNVs) have been convincingly associated with some neuropsychiatric disorders. From the observational studies of children and adults with similar behavioural profiles and learning disabilities linked to rare syndromes, often with reciprocal deletions and duplications (Elsea & Girirajan, 2008; Montcel & Mendizabai, 1996; PEOPLES et al., 2000; Potocki et al., 2007; Van der Aa et al., 2009; Yobb et al., 2005), to the association of a minority of cases of autism(Cuscó et al., 2009; Freitag, 2007; Marshall et al., 2008; Weiss et al., 2008), schizophrenia(International Schizophrenia Consortium, 2008; Kirov et

al., 2007; Levinson et al., 2011; Stefansson et al., 2008; Walsh et al., 2008; Xu et al., 2008) and ADHD (Lesch et al., 2011; N. M. Williams et al., 2010) with a range of deletions and duplications to the developing literature in bipolar disorder(Grozeva et al., 2012; Malhotra et al., 2011; Shane E McCarthy et al., 2009; McQuillin et al., 2011; Priebe et al., 2011; D. Zhang et al., 2009a) and unipolar disorder(Degenhardt et al., 2012; Glessner et al., 2010; Rucker et al., 2011), the discovery of rare CNVs has been an intriguing addition to a field with a relative lack of replicated genetic associations.

We have observed in our own study instances of CNVs not only associated with affective disorders, but also schizophrenia, bipolar disorder, autism and even mental retardation. In this perhaps unexpected context the concept of association in genetic disorders deserves special attention. Association of a behavioural phenotype with a genetic variant in itself is only weak evidence of causation, largely because the conceptual gap between the two observations is so large. If one CNV is associated with many disorders it is logical to think that the association with CNV lies not with the disorder itself, but more with some intermediate phenotype, or with a predisposition to a broad range of disorders. Psychiatric disorders most associated with cognitive impairment (learning disability, autism) were first to be associated with rare CNVs. Then schizophrenia and ADHD, both diagnoses associated with cognitive impairment and reduced IQ were quickly found to be associated. However the affective disorders, which characteristically do not have a large element of comorbid cognitive impairment, remain, on the balance of evidence, to be conclusively associated. An intuitive conclusion from this is that rare structural variation is

not associated with neuropsychiatric diagnoses per se, but is associated with the neurocognitive impairments that predate the morbid state.

As a case in point, in their study of ADHD, Williams et al(N. M. Williams et al., 2010) were careful to point out that their case cohort was of lower intelligence than the normal population, and in a re-analysis of this sample, Langley et al.(Langley et al., 2011) concluded that whilst those children with ADHD and a CNV were clinically indistinguishable from those children with ADHD who did not have a CNV, those children who did have a CNV were significantly more likely to be intellectually impaired (IQ < 70). We see, in systematic studies of patients with unexplained learning disability, that the detection rate for rare and de novo genetic variation is 2-3 fold higher with microarray technology than for G-banded karyotyping (10-15% vs. 5% respectively)(de Vries et al., 2005; Engels et al., 2007; Fan et al., 2007; Friedman et al., 2006; Jaillard et al., 2010). It is therefore perhaps not surprising that we go on, with this higher resolution technology, to detect the same rare genetic variation in a smaller minority of patients diagnosed with disorders that are themselves associated with cognitive impairment. We did not have measures of intellectual ability in our cases, but have made the point that our screened control group were derived largely from students and staff at Kings College London. The differences we observe may underline this factor.

That the simple association of CNV with observed clinical disorder may be a simplistic view is further reinforced by the variable penetrance of the CNVs observed in this dataset. The case of the 22q11.2 deletion in our case cohort

emphasises this point. Whilst this case suffers from a chronic depressive illness and social anxiety, he does not display the overt physical manifestations of the variant, and does not have a history of cognitive impairment or psychosis. Indeed his IQ is normal (101). Even this large CNV, covering a gene rich region, demonstrates variable penetrance. This suggests that an interacting concert of protective and predisposing variants probably contributes to observed disease states in individuals, or the lack of them in individuals with large and rare CNVs.

Perhaps an increasingly pressing question in this context is to what extent rare variants are present in the 'normal' population? Social and economic status is robustly linked to psychiatric disorders(Y. Yu & Williams, 1999). Are those who come into contact with mental health services with apparently 'abnormal' psychiatric states, as they are currently defined, merely people who are socially disadvantaged in other ways? The answers to such questions, as ever, are not dichotomous, but this in itself suggests a level of complexity that geneticists, indeed the statistical methods that underlie it, are not yet able to fully tackle. That said, our current approach of associating disease state with genetic variant is a concise and preliminary approach to an emergently tricky problem. The issue of how to accurately associate phenotypic end-points with genetic states when the association has the potential to be confounded by so many variables in-between is bewildering in its complexity. This is partly tackled by using matched control samples.

Even if we were to understand the genetic nature of disease, the question remains of what we would seek to do with the knowledge? Whilst some advocate

a direct link between genes and various outcomes, the reality is probably more subtle, with ever-changing environment and our relationships with it of similar importance. With this in mind, and considering complex diseases in particular, it does not seem likely that a genetic footprint of an individual could, in the majority of instances, provide either any reassurance of a disease not developing, as well as any certainty that it would develop. Probabilities of development would need to be based on population-based measurements, which do not necessarily apply to individuals. Those probabilities in turn would rely on a detailed knowledge of not only how individual risk variants contribute to disease, but also how they interact with each, both genetically and epigenetically. Such a challenge of discovery is indeed formidable, but in the development of tools to interrogate the genome the field of genetics is taking the first steps into this uncharted domain.

7.6 Future Directions

Future studies in the genetics of depressive disorder will either use existing sample sets or endeavour to collect new ones. Analysis of CNVs is particularly slippery, and is probably best performed on contemporaneously collected and processed DNA samples derived from venous blood, genotyped on identical arrays and processed with an identical analysis pipeline.

Our own dataset may be used for pathway analysis, to look for an over-representation of genes in specific pathways potentially giving greater insight into the aetiology of depression. This may include stress hormone pathways,

monoaminergic synthesis and neurotransmission pathways, pathways regulating neuronal synthesis and dendritic outgrowth, as well as immunological pathways.

As already discussed, whilst it is clear that rare and de novo CNVs occur in some individuals with psychiatric disorder, it is difficult to aetiologically link the variants to disease seen in our cohort without additional evidence from family members. A proportion of the DeNT sample also have genotype data from affected siblings and other family members, in which we will follow up the putatively aetiological CNVs observed in probands in due course.

Next generation sequencing, whilst certainly a step forward in the analysis of genomic variation in general, may not be an accurate means of deriving copy number states without additional validation from arrays or other methods, since they rely on read counts of sequences that may vary due to other factors. Nonetheless its power at detecting point mutations and other small variants cannot be under-estimated, and we have started work on exome sequencing for a proportion of the samples used in this study.

7.7 Conclusion

It has been over 60 years since the seminal discovery of the basic structure of DNA(Watson & Crick, 1953), and over 10 years since the publication of the first draft human genome sequence(Lander et al., 2001; Venter et al., 2001). The field has attracted both considerable criticism and considerable praise, and large capital investments, especially in developed societies. The vast raft of association

studies that have linked variants with disease come with them the realisation that the relative risks for individual variants are usually only modest, especially in the complex diseases that confer so much morbidity and mortality throughout the world. This hints at the interwoven and interdependent genetic complexity facing geneticists working on complex human diseases today. Nonetheless the field has made significant advances, especially since the advent of genome-wide association studies, and the rapid technological advances in genome sequencing and bioinformatics are especially promising. Depressive illness is an enigmatic and complex disorder which has so far resisted efforts to probe its genetic predispositions, however the observation of rare copy number variants in our sample provides evidence that this level of genetic variation may be of aetiological relevance in a minority of cases.

Chapter 8. Appendices

"Mr. Osborne, may I be excused? My brain is full."

8.1 Introduction

Within this section I include less important elements of our methodology and analysis which, to be concise, have been moved to this appendix chapter. Contents are headed by chapter, under the same paragraph heading as their reference in the main text, with prefix '8.'.

8.2 Chapter 2

8.2.3.2.5.4 Calling CNVs with PennCNV

Commands settings for PennCNV: detect_cnv.pl -test -hmm hhall.hmm -pfb hhall.pfb -listfile <filelist> -logfile <logfile> -output <callfile>

8.2.3.3 R Script for Visualizing LRR/BAF By Chr/Sample

Example with one chromosome, for one sample

one=read.table("<path_to_sample_file>", sep="\t", header=T, na.strings="NaN")

bitmap(file="<path_to_image_file>",type="jpeg",width=10,height=5,res=300,poin
tsize=10)

par(mfrow=c(2,1),las=1)

plot(one$Position[one$Chr=="<chr_number>"],one$<sample_id>.Log.R.Ratio[on
e$Chr=="<chr_number>"],pch=20,ylim=c(-4,2),ylab="Log R Ratio",xlab="Chr
<chr_number> Position",col="red",cex=0.1)

title(paste("<sample_id>.Chr.<chr_number>"))

plot(one$Position[one$Chr=="9"],one$<sample_id>.B.Allele.Freq[one$Chr==<ch

r_number>],pch=20,ylim=c(0,1),ylab="B Allele Frequency",xlab="Chr

<chr_number> Position",col="blue",cex=0.1)

dev.off()

R Script for Visualising Data by Marker

one=read.table("<path_to_marker_data_file>",sep="\t")

bitmap(file="<path_to_image_file>",type="jpeg",width=5,height=5,res=300,point

size=10)

par(mfrow=c(1,1),las=1)

plot(one$V8,one$V7,pch=20,ylab="Norm R",xlab="Norm

Theta",col="red",cex=0.1)

title(paste("<marker_name>"))

dev.off()

8.2.3.5 Exclusion Coordinate List

The following table illustrates genomic regions from which we removed CNV calls.

Area	Region 1	Region 2
Immunoglobulin	chr22:20715572-21595082	N/A
Immunoglobulin	chr14:105065301-106352275	N/A
Immunoglobulin	chr2:88937989-89411302	N/A
Immunoglobulin	chr14:21159897-22090937	N/A
Centromere	chr1:121000001-128100000	N/A
Centromere	chr2:90900001-95800000	N/A
Centromere	chr3:89300001-93300000	N/A
Centromere	chr4:48600001-52500000	N/A

Centromere	chr5:45700001-50600000	N/A
Centromere	chr6:58300001-63500000	N/A
Centromere	chr7:57300001-61200000	N/A
Centromere	chr8:43100001-48200000	N/A
Centromere	chr9:46600001-60400000	N/A
Centromere	chr10:38700001-42200000	N/A
Centromere	chr11:51300001-56500000	N/A
Centromere	chr12:33100001-36600000	N/A
Centromere	chr13:13400001-18500000	N/A
Centromere	chr14:13500001-19200000	N/A
Centromere	chr15:14000001-18500000	N/A
Centromere	chr16:34300001-40800000	N/A
Centromere	chr17:22000001-23300000	N/A
Centromere	chr18:15300001-17400000	N/A
Centromere	chr19:26600001-30300000	N/A
Centromere	chr20:25600001-28500000	N/A
Centromere	chr21:9900001-13300000	N/A
Centromere	chr22:9500001-16400000	N/A
Telomere	chr1:1-500000	chr1:246749719-247249719
Telomere	chr2:1-500000	chr2:242451149-242951149
Telomere	chr3:1-500000	chr3:199001827-199501827
Telomere	chr4:1-500000	chr4:190773063-191273063
Telomere	chr5:1-500000	chr5:180357866-180857866
Telomere	chr6:1-500000	chr6:170399992-170899992
Telomere	chr7:1-500000	chr7:158321424-158821424
Telomere	chr8:1-500000	chr8:145774826-146274826
Telomere	chr9:1-500000	chr9:139773252-140273252
Telomere	chr10:1-500000	chr10:134874737-135374737
Telomere	chr11:1-500000	chr11:133952384-134452384
Telomere	chr12:1-500000	chr12:131849534-132349534
Telomere	chr13:1-500000	chr13:113642980-114142980
Telomere	chr14:1-500000	chr14:105868585-106368585
Telomere	chr15:1-500000	chr15:99838915-100338915
Telomere	chr16:1-500000	chr16:88327254-88827254
Telomere	chr17:1-500000	chr17:78274742-78774742
Telomere	chr18:1-500000	chr18:75617153-76117153
Telomere	chr19:1-500000	chr19:63311651-63811651
Telomere	chr20:1-500000	chr20:61935964-62435964
Telomere	chr21:1-500000	chr21:46444323-46944323
Telomere	chr22:1-500000	chr22:49191432-49691432

Table 8.1. Regions excluded from our CNV analysis. Coordinates are based on build hg18.

8.2.3.5 Scripts for Restricting Call Sets to Rare CNVs

These scripts assume the PennCNV package(at least version dated August 2009) is installed and the raw PennCNV call file <call_file> is present in the current directory.

##Restrict calls to those made with more than 9 markers and more than 100kb in length

filter_cnv.pl -numsnp 10 -length 100k <call_file> > <call_file_a>

##Create a list of CNV calls with more than 50% overlap with regions within 500kb of the centromere or telomere, or any region known to rearrange in immortalised cell line creation

scan_region.pl <call_file_a> <imm_centro_telo_exclusion_regions> -minqueryfrac 0.5 > <call_file_b>

##Remove the above CNVs

fgrep -v -f <call_file_b> <call_file_a> > <call_file_b>

##Remove all calls from samples not passing predefined QC thresholds

fgrep -f <qcd_samples_list> <call_file_b> > <call_file_c>

##Create a list of poorly performing samples, defined as those having more than 20 CNV calls.

awk '{print $5}' <call_file_c> | sort | uniq -c | sort -k1nr | awk ' $1 > 20 {print $2} ' > <list_of_samples_failing_NCNV20>

##Remove calls from these samples

fgrep -v -f <list_of_samples_failing_NCNV20> <call_file_c> > <call_file_d>

##Create a list of common copy number regions, defined as those occurring in more than 1% of the remaining sample

awk '{print $1}' <call_file_d> | sort | uniq -c | sort -k1nr | awk ' $1 >

<N_samples/100> {print $2} ' > <CNP_regions_to_exclude>

##Create a list of CNV calls with more than 50% overlap over these regions

scan_region.pl <call_file_d> <CNP_regions_to_exclude> -minqueryfrac 0.5 >

<call_file_e>

##Remove these calls

fgrep -v -f <call_file_e> <call_file_d> > <call_file_f>

##Create a list of CNV calls where more than 20% of the call falls over a region where the marker density is less than 1 in 200kb.

scan_region.pl <call_file_f> <regions_gtkbp_marker_free> -minqueryfrac 0.2 >

<call_file_g>

##Remove these calls

fgrep -f -v <call_file_f> <call_file_g> > <final_call_file>

8.2.4.1 A Comparison of Samples With and Without Rare CNVs

Scripts for annotating CNVs with genic, exonic, intronic and intergenic names, and counting variants by type. These scripts rely on the correct installation and functioning of the PennCNV perl scripts filter_cnv.pl and scan_region.pl.

Genome Location	CNV Type	Script
Whole Genome	All	cat <call_file> | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'
Whole Genome	Deletions	cat <call_file> | filter_cnv.pl stdin -type del | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'
Whole Genome	Duplications	cat <call_file> | filter_cnv.pl stdin -type dup | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'

Intergenic	All	scan_region.pl <call_file> <gene_annotation_file> -refgene | grep NOT_FOUND | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'
Intergenic	Deletions	scan_region.pl <call_file> <gene_annotation_file> -refgene | grep NOT_FOUND | filter_cnv.pl stdin -type del | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'
Intergenic	Duplications	scan_region.pl <call_file> <gene_annotation_file> -refgene | grep NOT_FOUND | filter_cnv.pl stdin -type dup | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'
Genic	All	scan_region.pl <call_file> <gene_annotation_file> -refgene | grep -v NOT_FOUND | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'
Genic	Deletions	scan_region.pl <call_file> <gene_annotation_file> -refgene | grep -v NOT_FOUND | filter_cnv.pl stdin -type del | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'
Genic	Duplications	scan_region.pl <call_file> <gene_annotation_file> -refgene | grep -v NOT_FOUND | filter_cnv.pl stdin -type dup | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'
Intronic	All	scan_region.pl <call_file> <intron_coordinate_file> -minqueryfrac 1 | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'
Intronic	Deletions	scan_region.pl <call_file> <intron_coordinate_file> -minqueryfrac 1 | filter_cnv.pl stdin -type del | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'
Intronic	Duplications	scan_region.pl <call_file> <intron_coordinate_file> -minqueryfrac 1 | filter_cnv.pl stdin -type dup | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'
Exonic	All	scan_region.pl <call_file> <gene_annotation_file> -refexon | grep -v NOT_FOUND | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'
Exonic	Deletions	scan_region.pl <call_file> <gene_annotation_file> -refexon | grep -v NOT_FOUND | filter_cnv.pl stdin -type del | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'
Exonic	Duplications	scan_region.pl <call_file> <gene_annotation_file> -refexon | grep -v

		NOT_FOUND | filter_cnv.pl stdin -type dup | awk '{print $5}' | sort | uniq | wc | awk '{print $1}'

Table 8.2. Scripts for restricting and counting numbers of samples with CNVs in different regions of the genome.

8.2.4.1.4 Analysis Within Intronic Regions

Low numbers of samples with intronic CNVs only reduced our power to detect an effect. We found no significant differences between the frequencies of samples with different CNV types in the three cohorts (Tables 8.3 and 8.4, Figure 8.1).

	Cases (N=2723)	**Screened Controls (N=348)**	**P value (Pearson's chi²)**	**Odds Ratio (95% CI)**
No. samples with an intronic deletion**	60 2.2%	8 2.3%	0.85 (2-sided Fisher's exact)	0.85 0.46 - 1.99
No. samples with an intronic duplication**	25 0.9%	2 0.5%	0.76 (2-sided Fisher's exact)	1.60 0.42 - NaN
No. samples with an intronic deletion or duplication**	85 3.1%	10 2.8%	1.00 (2-sided Fisher's exact)	1.09 0.57 - 2.09

Table 8.3. Frequency of samples with deletion, duplication and any CNV in cases and screened controls with Pearson's chi² statistic and odds ratios with 95% confidence intervals.

	Cases (N=2723)	WTCCC2 Controls (N=4828)	P value (Chi^2)	Odds Ratio (95% CI)
No. samples with an intronic deletion**	60 2.2%	117 2.4%	0.54	0.91 0.66 - 1.24
No. samples with an intronic duplication**	25 0.9%	40 0.8%	0.69	1.11 0.67 - 1.82
No. samples with an intronic deletion or duplication**	85 3.1%	156 3.2%	0.79	0.96 0.74 - 1.26

Table 8.4. Frequency of samples with deletion, duplication and any CNV in cases and Wellcome Trust controls with Pearson's chi^2 statistic and odds ratios with 95% confidence intervals.

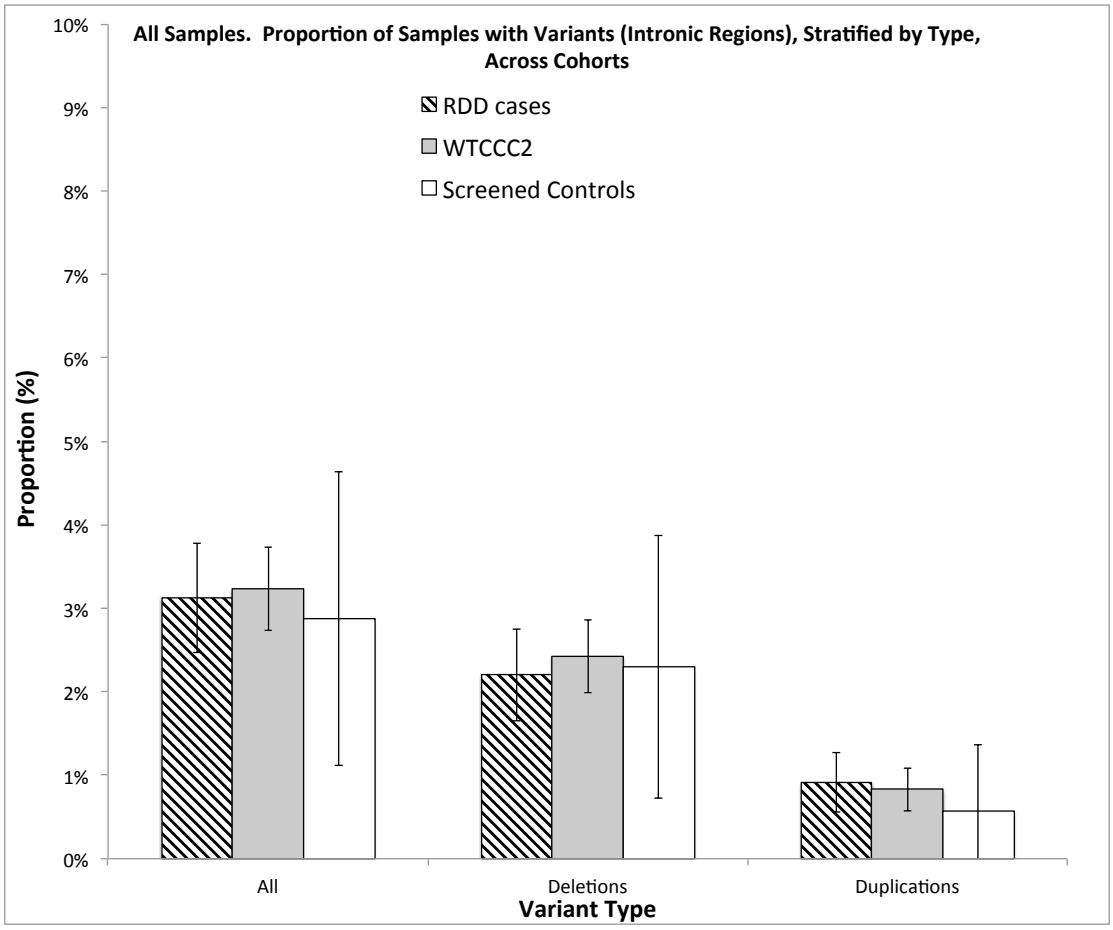


Fig 8.1. Proportion of samples with variants across only intronic regions of the genome, stratified by type, across cohorts.

8.2.4.1.5 Analysis Within Intergenic Regions

Within intergenic regions there were no significant differences in frequency of samples with CNVs of either type between the cases and screened controls. We found that our cases had significantly more CNVs than the WTCCC2 controls (p=0.0033, OR 1.17 (95% CI 1.05 - 1.30). This difference appeared to be predominantly driven by an increased frequency of samples with duplications in the RDD case cohort ($p=6.43x10^{-4}$, OR 1.29 (95% CI 1.11 - 1.49)) and, to a lesser extent, samples with deletions (p=0.068, OR 1.12 (95% CI 0.99 - 1.26)).

	Cases (N=2723)	**Screened Controls (N=348)**	**P value (Pearson's chi^2)**	**Odds Ratio (95% CI)**
No. samples with an intergenic deletion	523 19.2%	59 17.0%	0.31	1.16 0.87 - 1.56
No. samples with an intergenic duplication	345 12.7%	42 12.1%	0.75	1.06 0.75 - 1.49
No. samples with an intergenic deletion or duplication	796 29.3%	93 26.7%	0.31	1.13 0.88 - 1.46

Table 8.5. Frequency of samples with deletion, duplication and any CNV in cases and screened controls with Pearson's chi^2 statistic and odds ratios with 95% confidence intervals.

	Cases (N=2723)	WTCCC2 Controls (N=4828)	P value (Chi^2)	Odds Ratio (95% CI)
No. samples with an intergenic deletion	523 19.2%	846 17.5%	0.068	1.12 0.99 - 1.26
No. samples with an intergenic duplication	345 12.7%	488 10.1%	$6.43x10^{-4}$	1.29 1.11 - 1.49
No. samples with an intergenic deletion or duplication	796 29.3%	1260 26.1%	0.0033	1.17 1.05 - 1.30

Table 8.6. Frequency of samples with deletion, duplication and any CNV in cases and Wellcome Trust controls with Pearson's chi^2 statistic and odds ratios with 95% confidence intervals.

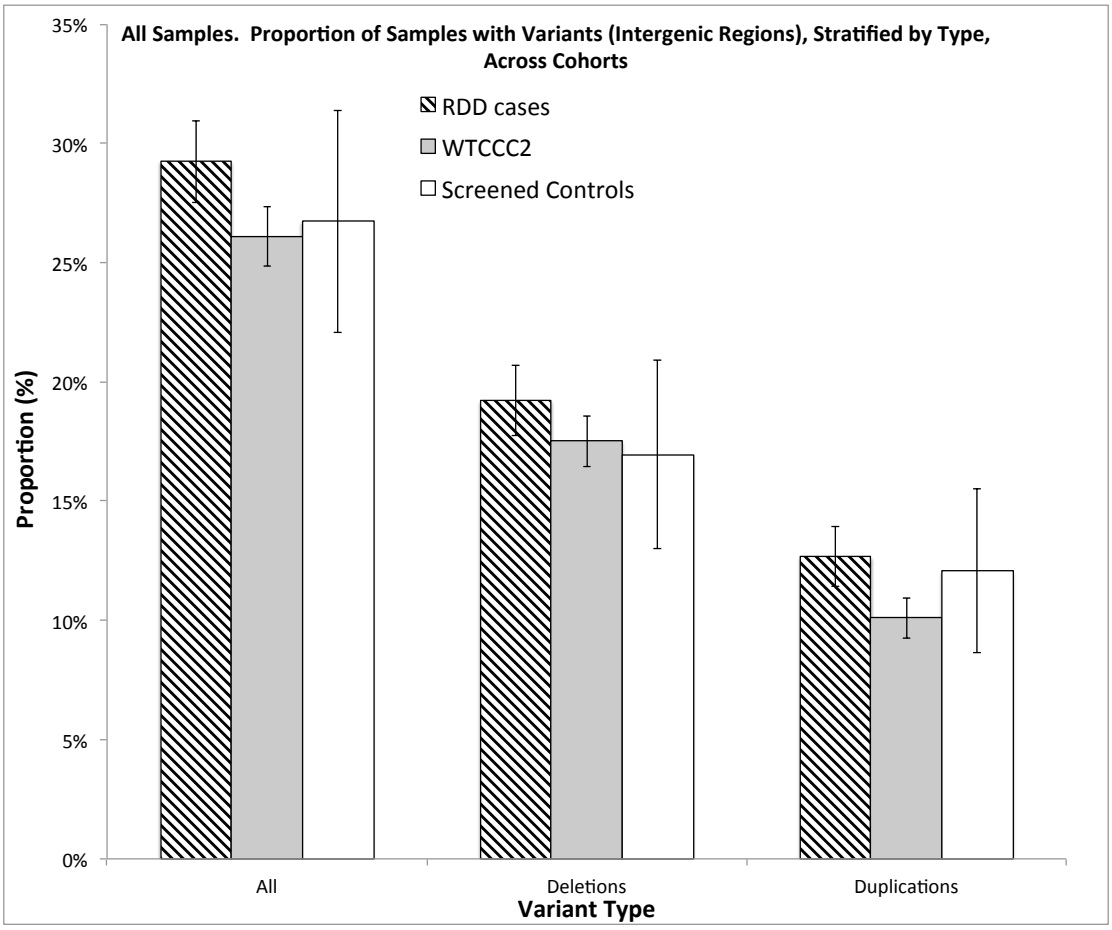


Fig 8.2. Proportion of samples with variants across intergenic regions of the genome, stratified by type, across cohorts.

8.2.4.7 Scripts for CNV Burden Analysis

plink --cfile <root_analysis_name> --cnv-indiv-perm --mperm 10000 </--cnv-dup/--cnv-del/> -out <root_analysis_name>

plink --cfile <root_analysis_name> --cnv-indiv-perm --mperm 10000 </--cnv-dup/--cnv-del/> --cnv-test-2sided -out <root_analysis_name>_2sided

8.2.4.8 Scripts for Singleton CNV Analysis

plink --cfile <cnv/fam/cnv.map_file_root_name> --cnv-indiv-perm --mperm 10000 --cnv-freq-exclude-above 1 (<--cnv-dup>/<cnv-del>)

8.3 Chapter 3

8.3.3.2 Genotyping and Quality Control

List of CNP regions with Illumina consensus marker set (610 vs modified 1M) tagSNPs.

CNP	Chr	CNV type	RSID	r Square
CNVR63.1	1	loss	rs12040542	1.000
CNVR92.2	1	gain	rs1010069	0.983
CNVR95.2	1	loss	rs2240335	0.895
CNVR102.1	1	loss	rs12139100	0.958
CNVR112.1	1	loss	rs7552167	0.986
CNVR138.1	1	loss	rs12132123	0.937
CNVR178.1	1	loss	rs7550236	1.000
CNVR187.3	1	gain	rs2117262	1.000
CNVR191.1	1	loss	rs595513	0.996
CNVR193.1	1	loss	rs12401924	0.872
CNVR200.1	1	loss	rs4915770	0.934
CNVR208.1	1	loss	rs3790426	0.997
CNVR216.1	1	loss	rs2815749	0.937
CNVR220.1	1	loss	rs4949623	0.948
CNVR221.1	1	loss	rs544606	1.000
CNVR227.1	1	loss	rs12065553	0.958

CNVR228.1	1	gain	rs1339118	1.000
CNVR238.1	1	loss	rs12568614	1.000
CNVR244.1	1	gain	rs12142922	0.905
CNVR267.1	1	gain	rs12143917	0.989
CNVR274.1	1	gain	rs6687976	0.954
CNVR276.1	1	loss	rs1710809	0.998
CNVR292.1	1	loss	rs10785835	0.996
CNVR312.1	1	loss	rs1290540	0.947
CNVR373.1	1	gain	rs1748383	0.990
CNVR374.1	1	gain	rs2779110	0.994
CNVR377.1	1	gain	rs9887904	0.947
CNVR388.1	1	loss	rs6426918	0.855
CNVR404.1	1	loss	rs16847563	1.000
CNVR422.1	1	gain	rs3001275	0.990
CNVR431.1	1	gain	rs7554573	1.000
CNVR432.1	1	loss	rs10912174	0.911
CNVR459.3	1	gain	rs6677604	0.987
CNVR463.1	1	loss	rs16844287	0.995
CNVR466.1	1	gain	rs2153279	0.990
CNVR472.1	1	gain	rs3087949	0.892
CNVR475.1	1	gain	rs3753036	0.986
CNVR506.1	1	loss	rs2646822	1.000
CNVR510.1	1	loss	rs1567602	0.988
CNVR532.1	1	loss	rs6700582	0.994
CNVR550.1	1	gain	rs16833645	0.950
CNVR560.1	1	loss	rs12095464	1.000
CNVR561.1	1	loss	rs12035407	1.000
CNVR563.1	1	gain	rs10737874	0.989
CNVR619.1	2	gain	rs6743920	0.986
CNVR625.3	2	loss	rs4854428	0.989
CNVR660.1	2	loss	rs6722795	0.881
CNVR665.1	2	gain	rs12692305	0.994
CNVR733.1	2	loss	rs829590	0.981
CNVR739.2	2	gain	rs3845781	0.826
CNVR776.1	2	loss	rs6730278	0.940
CNVR777.2	2	loss	rs6751281	0.951
CNVR785.1	2	loss	rs13035544	0.994
CNVR789.1	2	gain	rs7562270	0.971
CNVR832.1	2	gain	rs11897583	0.958
CNVR847.1	2	loss	rs7578883	0.997
CNVR886.1	2	loss	rs11691779	0.983
CNVR888.1	2	gain	rs11691779	0.971
CNVR888.2	2	gain	rs11691779	0.980
CNVR894.2	2	gain	rs4851439	1.000
CNVR896.1	2	loss	rs13029804	1.000
CNVR901.1	2	gain	rs12465089	0.989

CNVR905.1	2	loss	rs11124097	1.000
CNVR918.2	2	loss	rs17035400	0.994
CNVR943.1	2	gain	rs13019645	1.000
CNVR957.1	2	gain	rs4848220	1.000
CNVR963.1	2	loss	rs13013829	0.942
CNVR965.1	2	loss	rs1365783	1.000
CNVR966.1	2	loss	rs2114601	1.000
CNVR985.1	2	loss	rs7419565	0.995
CNVR1013.1	2	loss	rs6725195	0.997
CNVR1041.2	2	gain	rs2358143	1.000
CNVR1053.1	2	gain	rs2306797	1.000
CNVR1071.1	2	loss	rs13425972	0.976
CNVR1120.1	2	gain	rs10176518	0.970
CNVR1123.1	2	gain	rs13010104	0.997
CNVR1158.1	2	gain	rs2194549	1.000
CNVR1184.1	2	loss	rs11883732	0.931
CNVR1201.1	2	gain	rs1510510	1.000
CNVR1254.1	3	gain	rs17193911	0.981
CNVR1267.1	3	gain	rs7628494	0.996
CNVR1269.1	3	loss	rs9880886	0.928
CNVR1275.1	3	loss	rs4686115	0.989
CNVR1288.1	3	loss	rs17776719	0.864
CNVR1338.1	3	loss	rs9865001	0.913
CNVR1348.1	3	loss	rs9848822	0.993
CNVR1358.1	3	loss	rs9821993	1.000
CNVR1414.1	3	loss	rs2888253	0.976
CNVR1441.1	3	loss	rs6548795	0.956
CNVR1470.1	3	loss	rs6800438	1.000
CNVR1471.1	3	loss	rs17217760	0.997
CNVR1474.1	3	loss	rs13065665	0.916
CNVR1478.1	3	loss	rs1165925	0.987
CNVR1543.1	3	loss	rs1320900	0.998
CNVR1572.1	3	loss	rs6440299	1.000
CNVR1573.1	3	gain	rs12487322	0.957
CNVR1574.1	3	loss	rs10470458	1.000
CNVR1576.1	3	gain	rs1511559	1.000
CNVR1591.1	3	loss	rs11708859	1.000
CNVR1608.2	3	gain	rs206276	1.000
CNVR1638.1	3	loss	rs16825525	0.952
CNVR1642.1	3	loss	rs9843890	0.997
CNVR1648.2	3	gain	rs4624543	1.000
CNVR1675.1	3	gain	rs9827142	0.989
CNVR1677.1	3	loss	rs1844850	1.000
CNVR1691.1	3	gain	rs1675932	0.976
CNVR1699.2	3	loss	rs6794037	0.916
CNVR1756.1	4	loss	rs12501769	0.985

CNVR1790.2	4	loss	rs4689024	0.819
CNVR1805.1	4	gain	rs4689810	0.851
CNVR1809.1	4	gain	rs4235269	1.000
CNVR1820.1	4	loss	rs4302456	0.994
CNVR1834.1	4	loss	rs10002171	0.991
CNVR1841.1	4	loss	rs4406013	0.993
CNVR1844.1	4	loss	rs10011670	0.998
CNVR1844.2	4	loss	rs10011670	0.998
CNVR1859.1	4	loss	rs1510702	1.000
CNVR1868.1	4	loss	rs6554032	0.864
CNVR1877.2	4	loss	rs7657131	0.992
CNVR1897.1	4	loss	rs2345904	1.000
CNVR1899.1	4	loss	rs7659996	0.985
CNVR1900.1	4	loss	rs4323141	0.992
CNVR1913.1	4	gain	rs9998449	0.937
CNVR1922.1	4	loss	rs4865140	1.000
CNVR1935.1	4	gain	rs1455706	0.887
CNVR1941.1	4	loss	rs11728914	0.915
CNVR1945.1	4	gain	rs10002424	0.985
CNVR1957.2	4	gain	rs12513058	0.956
CNVR1959.3	4	loss	rs4148307	1.000
CNVR2008.1	4	loss	rs10516969	0.989
CNVR2013.2	4	gain	rs2136080	0.918
CNVR2046.1	4	loss	rs17576773	0.962
CNVR2056.1	4	loss	rs2204269	1.000
CNVR2076.2	4	loss	rs10021585	0.991
CNVR2080.2	4	loss	rs4388143	0.988
CNVR2108.2	4	loss	rs11100904	0.998
CNVR2119.2	4	gain	rs6835546	1.000
CNVR2126.1	4	loss	rs7682716	0.978
CNVR2138.1	4	gain	rs4501178	0.984
CNVR2153.1	4	loss	rs949794	0.989
CNVR2174.1	4	gain	rs13142152	0.997
CNVR2179.1	4	gain	rs6553951	0.992
CNVR2184.1	4	loss	rs13149928	0.929
CNVR2187.1	4	gain	rs2705988	0.991
CNVR2194.1	4	loss	rs6851088	0.998
CNVR2207.1	4	loss	rs13113415	0.883
CNVR2227.1	4	loss	rs6829485	1.000
CNVR2246.2	4	loss	rs13119341	0.998
CNVR2309.1	5	loss	rs10061160	0.981
CNVR2338.1	5	gain	rs10037687	0.858
CNVR2352.1	5	gain	rs886527	0.980
CNVR2370.1	5	loss	rs17602477	1.000
CNVR2395.1	5	loss	rs11746426	0.998
CNVR2400.1	5	loss	rs12516067	0.963

CNVR2439.1	5	loss	rs4072686	0.983
CNVR2445.2	5	loss	rs4590183	0.992
CNVR2457.1	5	loss	rs7715136	1.000
CNVR2458.1	5	loss	rs4288087	0.953
CNVR2471.1	5	loss	rs6865094	0.986
CNVR2473.1	5	loss	rs1010757	1.000
CNVR2483.2	5	gain	rs12522846	1.000
CNVR2520.1	5	loss	rs11746597	1.000
CNVR2523.1	5	loss	rs7700867	1.000
CNVR2530.1	5	loss	rs155060	0.997
CNVR2538.2	5	loss	rs6594474	1.000
CNVR2577.1	5	loss	rs1972623	1.000
CNVR2589.1	5	loss	rs3853394	0.979
CNVR2594.1	5	loss	rs6595594	0.945
CNVR2613.1	5	gain	rs7380956	0.994
CNVR2632.1	5	gain	rs7733492	0.997
CNVR2648.1	5	loss	rs11167544	0.996
CNVR2650.1	5	gain	rs9918229	0.996
CNVR2652.1	5	loss	rs10515653	1.000
CNVR2659.1	5	loss	rs4704970	1.000
CNVR2667.1	5	loss	rs445187	0.986
CNVR2668.1	5	loss	rs387661	0.976
CNVR2670.1	5	loss	rs1421862	0.991
CNVR2671.1	5	loss	rs17285868	0.804
CNVR2676.1	5	loss	rs384227	0.890
CNVR2684.1	5	loss	rs6897691	1.000
CNVR2711.1	5	gain	rs6870926	0.923
CNVR2744.3	6	loss	rs9405611	0.976
CNVR2753.1	6	loss	rs1011088	0.963
CNVR2762.1	6	loss	rs2765351	0.983
CNVR2769.1	6	loss	rs1106200	0.997
CNVR2785.1	6	gain	rs2038015	0.854
CNVR2789.1	6	gain	rs9370929	0.984
CNVR2793.2	6	loss	rs6930531	0.976
CNVR2801.2	6	loss	rs1322537	0.987
CNVR2803.1	6	loss	rs13208362	0.975
CNVR2841.6	6	loss	rs10484554	0.895
CNVR2847.1	6	gain	rs4713614	0.922
CNVR2847.2	6	gain	rs2395352	1.000
CNVR2850.1	6	loss	rs4713646	0.981
CNVR2880.1	6	gain	rs10948155	0.888
CNVR2885.1	6	gain	rs9369730	1.000
CNVR2899.1	6	loss	rs6916171	1.000
CNVR2916.1	6	loss	rs629506	1.000
CNVR2920.4	6	gain	rs4424070	1.000
CNVR2929.1	6	loss	rs2693074	0.997

CNVR2933.1	6	loss	rs213776	0.997
CNVR2937.1	6	loss	rs4501410	0.886
CNVR2938.1	6	loss	rs1552165	0.926
CNVR2940.1	6	loss	rs2064701	0.992
CNVR2961.1	6	gain	rs9293853	0.998
CNVR2962.1	6	gain	rs2349095	1.000
CNVR2965.1	6	gain	rs997765	0.996
CNVR2973.2	6	gain	rs9447791	0.962
CNVR2977.1	6	gain	rs818238	0.984
CNVR2987.1	6	loss	rs10944161	1.000
CNVR3014.1	6	loss	rs7760831	0.950
CNVR3022.1	6	loss	rs9322816	1.000
CNVR3025.1	6	loss	rs7769042	0.984
CNVR3057.1	6	loss	rs9490285	0.958
CNVR3074.1	6	gain	rs9373073	0.966
CNVR3075.1	6	loss	rs9389154	0.921
CNVR3076.1	6	loss	rs7749709	0.981
CNVR3083.2	6	gain	rs6570451	1.000
CNVR3092.1	6	loss	rs1970317	1.000
CNVR3107.1	6	loss	rs3778089	1.000
CNVR3112.1	6	loss	rs4870208	1.000
CNVR3126.1	6	loss	rs9365085	0.997
CNVR3150.1	6	loss	rs1325410	0.834
CNVR3206.1	6	loss	rs9366218	0.864
CNVR3239.1	7	gain	rs6463900	0.993
CNVR3241.1	7	loss	rs10270654	1.000
CNVR3258.3	7	loss	rs12155314	1.000
CNVR3289.1	7	gain	rs2215985	0.998
CNVR3294.1	7	loss	rs17166711	0.977
CNVR3312.1	7	gain	rs2074764	0.916
CNVR3335.1	7	loss	rs10245004	1.000
CNVR3342.1	7	loss	rs12701110	0.969
CNVR3356.1	7	gain	rs17171603	0.823
CNVR3390.1	7	loss	rs1525574	0.974
CNVR3394.1	7	loss	rs479006	1.000
CNVR3401.1	7	loss	rs10271211	0.975
CNVR3415.1	7	loss	rs2119445	0.994
CNVR3440.1	7	loss	rs887010	0.996
CNVR3441.1	7	gain	rs7802158	1.000
CNVR3448.2	7	loss	rs6460033	1.000
CNVR3475.1	7	loss	rs2722964	0.960
CNVR3477.1	7	loss	rs10245061	1.000
CNVR3495.1	7	gain	rs180267	0.995
CNVR3500.1	7	loss	rs17271261	0.996
CNVR3513.2	7	loss	rs6943474	0.994
CNVR3516.1	7	gain	rs370760	0.993

CNVR3527.2	7	gain	rs4730401	0.998
CNVR3551.1	7	loss	rs2214706	0.987
CNVR3581.1	7	loss	rs357372	0.973
CNVR3599.1	7	loss	rs10226541	0.943
CNVR3601.1	7	loss	rs12703672	0.975
CNVR3604.1	7	gain	rs2533084	0.973
CNVR3614.2	7	gain	rs6950421	0.997
CNVR3631.1	7	gain	rs747561	1.000
CNVR3661.1	7	gain	rs10235911	0.826
CNVR3730.1	8	gain	rs7462690	0.995
CNVR3745.1	8	gain	rs6998429	1.000
CNVR3746.1	8	gain	rs7008391	0.998
CNVR3747.2	8	gain	rs4990839	0.989
CNVR3747.1	8	gain	rs17326768	0.910
CNVR3769.2	8	loss	rs12680018	0.908
CNVR3781.2	8	loss	rs1560971	0.973
CNVR3802.3	8	gain	rs1721073	0.998
CNVR3814.1	8	loss	rs1441762	0.995
CNVR3815.1	8	loss	rs2410629	1.000
CNVR3831.1	8	gain	rs6996838	0.824
CNVR3843.1	8	gain	rs12335264	0.991
CNVR3844.1	8	loss	rs1002149	1.000
CNVR3862.1	8	loss	rs4236929	1.000
CNVR3863.1	8	gain	rs10808939	0.969
CNVR3882.1	8	loss	rs1903311	1.000
CNVR3887.1	8	loss	rs17253706	1.000
CNVR3891.2	8	loss	rs2376422	1.000
CNVR3892.1	8	gain	rs7387992	1.000
CNVR3898.4	8	gain	rs10504229	0.979
CNVR3898.3	8	gain	rs2318144	0.984
CNVR3910.1	8	loss	rs1376767	1.000
CNVR3913.1	8	loss	rs10104164	0.990
CNVR3928.1	8	loss	rs10095184	0.900
CNVR3936.2	8	loss	rs2570171	0.995
CNVR3946.1	8	gain	rs7839273	1.000
CNVR3961.1	8	loss	rs391916	0.998
CNVR3965.1	8	loss	rs7011436	0.922
CNVR3978.2	8	gain	rs6985568	1.000
CNVR3996.1	8	loss	rs16870788	0.882
CNVR4014.1	8	gain	rs10104559	1.000
CNVR4020.1	8	loss	rs7018251	1.000
CNVR4022.1	8	loss	rs16886291	1.000
CNVR4045.1	8	loss	rs6470643	1.000
CNVR4054.1	8	loss	rs7006137	1.000
CNVR4057.1	8	loss	rs7013213	0.995
CNVR4070.1	8	gain	rs4631498	0.995

CNVR4145.1	9	gain	rs10961910	0.997
CNVR4168.1	9	loss	rs1052489	0.984
CNVR4173.1	9	gain	rs10511491	0.998
CNVR4220.1	9	loss	rs12554150	1.000
CNVR4231.1	9	loss	rs6476179	0.908
CNVR4332.1	9	loss	rs1890639	0.997
CNVR4344.1	9	loss	rs10115058	0.802
CNVR4380.1	9	loss	rs1475524	1.000
CNVR4401.1	9	loss	rs12683791	1.000
CNVR4414.1	9	gain	rs2808522	1.000
CNVR4454.1	9	gain	rs894208	0.975
CNVR4456.1	9	loss	rs2174926	0.994
CNVR4486.1	9	gain	rs883342	0.991
CNVR4500.1	9	gain	rs11792737	0.837
CNVR4509.1	9	gain	rs491220	0.975
CNVR4510.1	9	gain	rs10993787	0.831
CNVR4565.1	10	gain	rs11599917	0.883
CNVR4590.5	10	loss	rs11251694	0.808
CNVR4599.1	10	loss	rs7086693	0.991
CNVR4615.3	10	gain	rs10905399	0.998
CNVR4616.1	10	loss	rs2184380	1.000
CNVR4631.1	10	loss	rs1953307	0.892
CNVR4647.1	10	loss	rs9417254	0.953
CNVR4658.1	10	loss	rs868691	0.985
CNVR4660.1	10	gain	rs12766630	0.997
CNVR4663.1	10	loss	rs2150157	1.000
CNVR4675.1	10	loss	rs1054767	1.000
CNVR4679.1	10	gain	rs1624116	0.996
CNVR4707.2	10	loss	rs11239139	0.998
CNVR4738.1	10	loss	rs1194501	0.873
CNVR4740.1	10	gain	rs11003159	1.000
CNVR4741.1	10	loss	rs10824859	0.997
CNVR4750.1	10	loss	rs7099707	0.990
CNVR4751.1	10	loss	rs1822760	0.826
CNVR4760.1	10	loss	rs4297386	0.958
CNVR4761.1	10	gain	rs10826108	0.973
CNVR4762.1	10	gain	rs2658596	1.000
CNVR4782.1	10	loss	rs12268529	0.804
CNVR4783.1	10	gain	rs2394280	0.976
CNVR4792.1	10	loss	rs5006438	1.000
CNVR4819.1	10	loss	rs2342606	0.952
CNVR4846.1	10	loss	rs10785914	0.996
CNVR4866.3	10	loss	rs3853516	0.992
CNVR4882.1	10	loss	rs7903001	1.000
CNVR4906.1	10	gain	rs10886694	1.000
CNVR4908.1	10	loss	rs11199704	0.984

CNVR4909.1	10	gain	rs11199966	0.982
CNVR4911.1	10	loss	rs3750847	0.988
CNVR4931.2	10	gain	rs7904713	0.996
CNVR4934.1	10	loss	rs1556659	0.987
CNVR4949.1	10	loss	rs1891787	0.994
CNVR4954.1	10	loss	rs4382815	0.980
CNVR5019.1	11	gain	rs3817197	0.998
CNVR5024.1	11	gain	rs886277	0.993
CNVR5031.2	11	gain	rs10834648	0.967
CNVR5033.1	11	loss	rs7131138	0.993
CNVR5039.1	11	loss	rs2595994	1.000
CNVR5044.1	11	loss	rs10500643	0.898
CNVR5048.1	11	gain	rs7928052	1.000
CNVR5049.1	11	gain	rs1453415	0.997
CNVR5073.1	11	loss	rs734462	1.000
CNVR5093.1	11	loss	rs10766561	0.954
CNVR5102.1	11	loss	rs7358426	0.978
CNVR5105.1	11	loss	rs11027318	1.000
CNVR5119.1	11	loss	rs1493663	0.916
CNVR5142.1	11	loss	rs10836919	0.945
CNVR5151.1	11	loss	rs2862385	0.991
CNVR5154.1	11	gain	rs11606984	0.890
CNVR5158.3	11	gain	rs11038440	0.990
CNVR5162.1	11	gain	rs2291443	0.980
CNVR5177.2	11	gain	rs1391576	0.994
CNVR5178.2	11	gain	rs11229374	0.968
CNVR5186.1	11	loss	rs11229411	0.995
CNVR5187.1	11	loss	rs566084	0.992
CNVR5188.2	11	loss	rs1938725	0.928
CNVR5231.1	11	loss	rs3741132	0.947
CNVR5261.1	11	loss	rs3793975	0.988
CNVR5272.2	11	loss	rs7946764	1.000
CNVR5275.1	11	gain	rs2000960	0.943
CNVR5278.1	11	loss	rs6483491	1.000
CNVR5279.1	11	loss	rs1452939	0.995
CNVR5285.2	11	gain	rs658789	0.991
CNVR5292.2	11	loss	rs11226106	0.888
CNVR5293.1	11	gain	rs924561	1.000
CNVR5301.1	11	loss	rs2187036	0.990
CNVR5303.2	11	loss	rs9787772	0.854
CNVR5305.1	11	loss	rs10466556	0.994
CNVR5314.1	11	loss	rs10891422	0.971
CNVR5317.1	11	loss	rs1455650	0.995
CNVR5326.1	11	loss	rs1073635	0.852
CNVR5334.1	11	gain	rs540029	0.997
CNVR5344.1	11	loss	rs684534	0.979

CNVR5351.1	11	gain	rs549809	0.998
CNVR5386.1	12	gain	rs11064562	0.876
CNVR5432.1	12	gain	rs10845279	0.998
CNVR5439.1	12	loss	rs10845623	1.000
CNVR5459.1	12	loss	rs11486	0.951
CNVR5471.1	12	loss	rs12230192	1.000
CNVR5479.1	12	loss	rs2564577	0.991
CNVR5481.1	12	loss	rs12580209	0.917
CNVR5490.1	12	loss	rs1500072	1.000
CNVR5508.1	12	loss	rs11176706	0.901
CNVR5556.2	12	gain	rs4760288	0.997
CNVR5557.1	12	loss	rs6581261	0.909
CNVR5564.1	12	loss	rs1583845	0.991
CNVR5585.1	12	loss	rs7956954	0.986
CNVR5591.1	12	loss	rs10748210	0.978
CNVR5592.1	12	gain	rs1371580	0.997
CNVR5604.1	12	loss	rs11114991	0.950
CNVR5616.1	12	loss	rs4842509	0.984
CNVR5639.1	12	loss	rs7960340	0.959
CNVR5663.1	12	loss	rs12297204	0.969
CNVR5688.1	12	gain	rs7314743	0.988
CNVR5692.1	12	loss	rs10507257	0.995
CNVR5696.1	12	loss	rs2221165	0.906
CNVR5749.1	12	loss	rs3864899	0.994
CNVR5848.2	13	loss	rs2273893	0.995
CNVR5849.1	13	loss	rs4941870	0.998
CNVR5852.1	13	loss	rs2324165	1.000
CNVR5854.1	13	loss	rs9315681	0.997
CNVR5858.1	13	loss	rs17536002	1.000
CNVR5882.1	13	gain	rs9536723	0.901
CNVR5888.2	13	loss	rs500574	0.979
CNVR5897.1	13	loss	rs2770909	1.000
CNVR5909.2	13	loss	rs2324909	0.996
CNVR5910.1	13	loss	rs9541152	0.905
CNVR5955.1	13	loss	rs12868420	1.000
CNVR5971.1	13	gain	rs7328769	1.000
CNVR5985.1	13	gain	rs9517368	0.992
CNVR6004.3	13	gain	rs9587457	0.982
CNVR6005.1	13	loss	rs2766094	0.860
CNVR6060.2	13	loss	rs9604573	0.994
CNVR6079.1	14	loss	rs4981390	0.989
CNVR6097.2	14	gain	rs8018861	0.993
CNVR6098.1	14	loss	rs1108625	1.000
CNVR6103.1	14	loss	rs996912	1.000
CNVR6117.1	14	loss	rs11156875	1.000
CNVR6123.1	14	loss	rs1766132	0.979

CNVR6125.1	14	loss	rs10137084	0.991
CNVR6159.1	14	loss	rs1188157	0.998
CNVR6167.1	14	loss	rs7152532	0.987
CNVR6174.1	14	loss	rs10873180	1.000
CNVR6209.2	14	loss	rs3844534	0.987
CNVR6218.1	14	gain	rs1481405	1.000
CNVR6244.1	14	loss	rs8015373	1.000
CNVR6248.1	14	gain	rs11625680	0.982
CNVR6307.1	15	gain	rs8041086	0.993
CNVR6310.3	15	gain	rs1587351	0.967
CNVR6349.1	15	loss	rs1978498	0.836
CNVR6355.1	15	loss	rs4923698	1.000
CNVR6378.1	15	loss	rs2463386	0.993
CNVR6384.1	15	loss	rs11635005	0.997
CNVR6395.1	15	loss	rs17551084	0.976
CNVR6412.1	15	loss	rs7180768	1.000
CNVR6430.1	15	loss	rs11631249	0.990
CNVR6439.2	15	gain	rs11635147	0.962
CNVR6466.1	15	loss	rs1865366	0.994
CNVR6484.2	15	loss	rs10444974	0.985
CNVR6495.1	15	loss	rs9635398	1.000
CNVR6501.1	15	loss	rs12442026	0.924
CNVR6511.1	15	loss	rs719987	0.997
CNVR6531.3	15	loss	rs12900588	0.996
CNVR6536.1	15	gain	rs1867157	0.998
CNVR6540.1	15	gain	rs2654987	0.993
CNVR6622.1	16	gain	rs11077219	0.997
CNVR6623.1	16	loss	rs1567144	0.993
CNVR6664.1	16	loss	rs12446632	1.000
CNVR6674.1	16	loss	rs381901	0.881
CNVR6719.1	16	loss	rs4785259	0.943
CNVR6730.1	16	loss	rs2244613	0.829
CNVR6731.1	16	gain	rs2244613	0.816
CNVR6772.1	16	loss	rs1423730	0.993
CNVR6773.1	16	loss	rs12597434	0.932
CNVR6786.1	16	loss	rs12921650	0.997
CNVR6791.3	16	loss	rs4624193	0.996
CNVR6796.1	16	loss	rs2656646	0.910
CNVR6810.1	16	gain	rs12921771	0.993
CNVR6825.1	16	loss	rs416563	1.000
CNVR6837.1	16	gain	rs16940186	0.965
CNVR6843.1	16	loss	rs11640888	0.906
CNVR6983.1	17	gain	rs8067532	0.985
CNVR6999.1	17	loss	rs12940030	1.000
CNVR7043.1	17	loss	rs2199750	1.000
CNVR7071.1	17	loss	rs317400	0.976

CNVR7073.1	17	loss	rs1036495	0.969
CNVR7074.1	17	loss	rs7220773	0.985
CNVR7096.1	17	loss	rs2191377	1.000
CNVR7097.1	17	loss	rs10512485	0.967
CNVR7125.2	17	loss	rs9898058	0.996
CNVR7129.1	17	loss	rs9915136	0.994
CNVR7145.1	17	gain	rs2586094	0.997
CNVR7181.1	17	loss	rs974654	0.992
CNVR7226.3	17	loss	rs12943041	0.924
CNVR7241.1	18	gain	rs1940954	0.949
CNVR7243.1	18	loss	rs408299	0.939
CNVR7260.1	18	loss	rs7235783	0.901
CNVR7283.1	18	gain	rs1436904	1.000
CNVR7329.1	18	loss	rs2292382	0.967
CNVR7364.1	18	loss	rs10503105	0.864
CNVR7375.1	18	loss	rs3819218	0.993
CNVR7387.2	18	gain	rs12969439	1.000
CNVR7428.1	18	loss	rs9958999	0.991
CNVR7492.2	19	gain	rs11084971	0.934
CNVR7512.1	19	gain	rs12983349	0.890
CNVR7573.2	19	gain	rs1579445	0.967
CNVR7581.2	19	gain	rs390911	1.000
CNVR7584.1	19	loss	rs1658216	0.984
CNVR7627.1	19	loss	rs4806152	0.888
CNVR7645.2	19	gain	rs305954	0.837
CNVR7706.1	19	loss	rs324121	1.000
CNVR7738.1	19	gain	rs2018863	0.997
CNVR7762.1	20	gain	rs8124053	0.996
CNVR7763.2	20	loss	rs4814391	0.963
CNVR7794.1	20	loss	rs7270966	1.000
CNVR7796.2	20	gain	rs2143553	0.899
CNVR7798.2	20	gain	rs8120353	0.834
CNVR7853.2	20	loss	rs8121830	0.907
CNVR7875.1	20	loss	rs2296239	0.997
CNVR7876.2	20	gain	rs6014120	0.887
CNVR7881.1	20	loss	rs1361051	0.998
CNVR7956.1	21	gain	rs13046557	0.986
CNVR8007.1	21	loss	rs2836565	0.923
CNVR8105.1	22	loss	rs2283802	0.954
CNVR8140.1	22	gain	rs5998902	1.000

Table 8.7. List of CNP tagSNPs used in our analysis with r^2 values.

8.3.3.3 Association Testing Scripts

plink --bfile <bed/bim/fam root file name> --pheno <phenotype file> --covar <covariates file> --covar-name e1,e2 --snps <snp list> --logistic --ci 0.95 --out <output root file name> --perm

8.4 Chapter 4

8.4.3.4.1 PLINK Scripts for PennCNV Data

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-indiv-perm --mperm 10000 --out <root_file_name>_allcnv_allvar_mperm

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-indiv-perm --mperm 10000 --cnv-del --out <root_file_name>_allcnv_del_mperm

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-indiv-perm --mperm 10000 --cnv-dup --out <root_file_name>_allcnv_dup_mperm

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-freq-exclude-below <1%_QC_sample_N> --cnv-overlap 0.5 --cnv-indiv-perm --mperm 10000 --out <root_file_name>_commoncnv_allvar_mperm

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-freq-exclude-below <1%_QC_sample_N> --cnv-overlap 0.5 --cnv-indiv-perm --mperm 10000 --cnv-del --out <root_file_name>_commoncnv_del_mperm

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-freq-exclude-below <1%_QC_sample_N> --cnv-overlap 0.5 --cnv-indiv-perm --mperm 10000 --cnv-dup --out <root_file_name>_commoncnv_dup_mperm

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-freq-exclude-above <1%_QC_sample_N> --cnv-overlap 0.5 --cnv-indiv-perm --mperm 10000 --out <root_file_name>_rarecnv_allvar_mvperm

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-freq-exclude-above <1%_QC_sample_N> --cnv-overlap 0.5 --cnv-indiv-perm --mperm 10000 --cnv-del --out <root_file_name>_rarecnv_del_mperm

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-freq-exclude-above <1%_QC_sample_N> --cnv-overlap 0.5 --cnv-indiv-perm --mperm 10000 --cnv-dup --out <root_file_name>_rarecnv_dup_mperm

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-indiv-perm --mperm 10000 --out <root_file_name>_allcnv_allvar_2sidemperm --cnv-test-2sided

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-indiv-perm --mperm 10000 --cnv-del --out <root_file_name>_allcnv_del_2sidemperm --cnv-test-2sided

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-indiv-perm --mperm 10000 --cnv-dup --out <root_file_name>_allcnv_dup_2sidemperm --cnv-test-2sided

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-freq-exclude-below <1%_QC_sample_N> --cnv-overlap 0.5 --cnv-indiv-perm --mperm 10000 --out <root_file_name>_commoncnv_allvar_2sidemperm --cnv-test-2sided

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-freq-exclude-below <1%_QC_sample_N> --cnv-overlap 0.5 --cnv-indiv-perm --mperm 10000 --cnv-del --out <root_file_name>_commoncnv_del_2sidemperm --cnv-test-2sided

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-freq-exclude-below <1%_QC_sample_N> --cnv-overlap 0.5 --cnv-indiv-perm --mperm 10000 --cnv-dup --out <root_file_name>_commoncnv_dup_2sidemperm --cnv-test-2sided

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-freq-exclude-above <1%_QC_sample_N> --cnv-overlap 0.5 --cnv-indiv-perm --mperm 10000 --out <root_file_name>_rarecnv_allvar_2sidemperm --cnv-test-2sided

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-freq-exclude-above <1%_QC_sample_N> --cnv-overlap 0.5 --cnv-indiv-perm --mperm 10000 --cnv-del --out <root_file_name>_rarecnv_del_2sidemperm --cnv-test-2sided

plink --cnv-list <cnv_file> --fam <fam_file> --map plink.cnv.map --cnv-freq-exclude-above <1%_QC_sample_N> --cnv-overlap 0.5 --cnv-indiv-perm --mperm 10000 --cnv-dup --out <root_file_name>_rarecnv_dup_2sidemperm --cnv-test-2sided

8.4.3.4.2 DNAcopy R Scripts

We used the following commands to process a sample of the name <sample_id> within DNAcopy and repeated this function for all samples

<sample_id><-read.table("<sample_id>_logRatio_for_DNA_copy.txt", sep="\t", header=TRUE, as.is=TRUE)

<sample_id>_cna <- CNA(<sample_id>$<sample_id>,<sample_id>$Chromosome, <sample_id>$Position, data.type="logratio", sampleid="<sample_id>")

<sample_id>_cna_smoothed <- smooth.CNA(<sample_id>_cna)

<sample_id>_cna_smoothed_segmented <- segment(<sample_id>_cna_smoothed, undo.splits="sdundo", undo.SD = 3, verbose = 1)

<sample_id>_cna_smoothed_segmented_table <- print(<sample_id>_cna_smoothed_segmented)

write.table(<sample_id>_cna_smoothed_segmented_table,file="<sample_id>_cna_smoothed_segmented.txt",sep="\t")

bitmap(file="<sample_id>_dna_copy.pdf",type="pdfwrite",width=10,height=10,res=300)

plot(<sample_id>_cna_smoothed_segmented, plot.type="s")

dev.off()

<sample_id>_cna_smoothed_segmented_pvalues<- segments.p(<sample_id>_cna_smoothed_segmented)

write.table(<sample_id>_cna_smoothed_segmented_pvalues,sep="\t",file="<sample_id>_cna_smoothed_segmented_pvalues.txt")

<sample_id>_cna_smoothed_segmented_summary<-segments.summary(<sample_id>_cna_smoothed_segmented)

write.table(<sample_id>_cna_smoothed_segmented_summary,sep="\t",file="<sample_id>_cna_smoothed_segmented_summary.txt")

bitmap(file="<sample_id>_chr22_dna_copy.pdf",type="pdfwrite",width=10,height=5,res=300)

plotSample(<sample_id>_cna_smoothed_segmented, sampleid=1, chromlist=22, main="<sample_id>_Chr_22", xlab="Probe_index")

dev.off()

8.4.4.2 Follow Up of 22q11.2 PennCNV Calls with array CGH

Following is a complete table of CNVs followed up by array CGH in the 22q11.2 region.

Location	NumSNP	Length	Type	Sample	Validates?
chr22:22676385-22717669	6	41,285	Dup	B0095H7	Yes
chr22:17257787-17388108	41	130,322	Del	B0095H6	Yes
chr22:22676385-22717669	6	41,285	Dup	B0095E8	Yes
chr22:22676385-22717669	6	41,285	Del	B0095D7	Yes
chr22:20645312-20903637	190	258,326	Del	B0095D5	Yes
chr22:22676385-22717669	6	41,285	Dup	B0095D0	Yes
chr22:23983992-24244593	69	260,602	Del	B0095C7	Yes
chr22:23991725-23999581	5	7,857	Dup	B0095AZ	Yes
chr22:22676385-22717669	6	41,285	Dup	B0095AL	Yes
chr22:22676385-22717669	6	41,285	Dup	B00959R	Yes
chr22:22676385-22717669	6	41,285	Dup	B00959K	Yes
chr22:23991725-24240667	67	248,943	Dup	B00959I	Yes
chr22:22676385-22717669	6	41,285	Dup	B0090V3	Yes
chr22:23994408-24162419	33	168,012	Del	B0090V0	Yes

chr22:24202592-24240667	25	38,076	Del	B0090V0	Yes
chr22:22619365-22653131	5	33,767	Dup	B0090UX	No
chr22:22664948-22698161	6	33,214	Del	B0090UX	Yes
chr22:23991725-24240667	67	248,943	Del	B0090UQ	Yes
chr22:23994408-24244593	67	250,186	Del	B0090UF	Yes
chr22:22038020-23327473	307	1,289,454	Dup	B0090TI	Yes
chr22:17257787-18621160	362	1,363,374	Dup	B0090SV	Yes
chr22:22676385-22717669	6	41,285	Dup	B0090R2	Yes
chr22:23994408-24240667	66	246,260	Del	B0090QK	Yes
chr22:22664948-22698161	6	33,214	Del	B0090Q1	Yes
chr22:22676385-22717669	6	41,285	Del	B0090KT	Yes
chr22:17257787-17388108	41	130,322	Del	B008CBL	Yes
chr22:21361579-21370362	5	8,784	Dup	B008CBL	No
chr22:22619365-22653131	5	33,767	Dup	B008CBJ	No
chr22:22664948-22698161	6	33,214	Del	B008CBJ	Yes
chr22:22664948-22698161	6	33,214	Del	B008CBC	Yes
chr22:22676385-22717669	6	41,285	Dup	B008CAX	Yes
chr22:22676385-22728586	7	52,202	Dup	B008CA9	Yes
chr22:23983992-24244593	69	260,602	Del	B008CA5	Yes
chr22:22619365-22653131	5	33,767	Dup	B008C7E	No
chr22:22664948-22698161	6	33,214	Del	B008C7E	Yes
chr22:23994408-24162419	33	168,012	Del	B008C77	Yes
chr22:24219385-24244593	14	25,209	Del	B008C77	Yes
chr22:22619365-22653131	5	33,767	Dup	B008C4Y	No
chr22:22664948-22698161	6	33,214	Del	B008C4Y	Yes
chr22:23999142-24240667	65	241,526	Del	B008C47	Yes
chr22:20645312-20901854	189	256,543	Dup	B008C30	Yes
chr22:17269490-17398986	42	129,497	Dup	B008C2H	Yes
chr22:22676385-22717669	6	41,285	Del	B008C2H	Yes
chr22:20637381-20903637	191	266,257	Dup	B008BZU	Yes
chr22:21041762-21064671	22	22,910	Dup	B008BZU	Yes
chr22:21361579-21419339	18	57,761	Dup	B008BZU	No
chr22:22664948-22717669	9	52,722	Dup	B008BZU	Yes
chr22:23994408-24230823	59	236,416	Del	B008BZK	Yes
chr22:17281004-17355587	26	74,584	Del	B008BZ2	No
chr22:18151568-18185263	14	33,696	Del	B008BZ2	No
chr22:21769530-21802954	11	33,425	Del	B008BZ2	No
chr22:17241748-17388108	43	146,361	Del	B008BXP	Yes
chr22:22059698-22077079	6	17,382	Del	B008BVP	No
chr22:17426677-17578226	40	151,550	Dup	B008BVB	Yes
chr22:22619365-22653131	5	33,767	Dup	B008BVB	No
chr22:22664948-22698161	6	33,214	Del	B008BVB	Yes
chr22:21441861-21554058	11	112,198	Del	B008BUZ	Yes
chr22:22664948-22698161	6	33,214	Del	B008BU5	Yes
chr22:18283733-18317638	19	33,906	Del	B008BU4	No
chr22:18328337-18332561	8	4,225	Del	B008BU4	No
chr22:18334458-18347543	9	13,086	Del	B008BU4	No
chr22:18596156-18638973	11	42,818	Del	B008BU4	No
chr22:19674884-19761181	49	86,298	Del	B008BU4	No
chr22:20312892-20327070	5	14,179	Del	B008BU4	No
chr22:22059698-22123801	21	64,104	Del	B008BU4	No
chr22:23994408-24240667	66	246,260	Del	B008BTU	Yes
chr22:22676385-22717669	6	41,285	Del	B008BT0	Yes
chr22:23994408-24244593	67	250,186	Dup	B008BQ6	Yes
chr22:22676385-22717669	6	41,285	Dup	B008BQ5	Yes
chr22:22619365-22664948	6	45,584	Dup	B008BPP	No
chr22:21361579-21370362	5	8,784	Dup	B008BPM	No
chr22:22619365-22653131	5	33,767	Del	B008BPL	No
chr22:22664948-22717669	9	52,722	Dup	B008BPL	Yes
chr22:22619365-22653131	5	33,767	Del	B008BPK	No
chr22:22664948-22698161	6	33,214	Dup	B008BPK	Yes

chr22:23991725-24240667	67	248,943	Del	B008BPK	Yes
chr22:22664948-22717669	9	52,722	Dup	B008BPI	Yes
chr22:22676385-22717669	6	41,285	Del	B008BOZ	Yes
chr22:23983992-24244593	69	260,602	Del	B008BOW	Yes
chr22:22676385-22717669	6	41,285	Del	B008BMN	Yes
chr22:22145573-22158684	7	13,112	Del	B008BM5	No
chr22:23994408-24240667	66	246,260	Del	B008BM5	Yes
chr22:23994408-24244593	67	250,186	Del	B008BLS	Yes
chr22:23999142-24240667	65	241,526	Del	B007WED	Yes
chr22:23983992-24128856	30	144,865	Dup	B007WCU	Yes
chr22:24202592-24238441	23	35,850	Dup	B007WCU	Yes
chr22:22676385-22717669	6	41,285	Dup	B007W9X	Yes
chr22:20679088-20903637	182	224,550	Del	B007VGQ	Yes
chr22:23994408-24239811	65	245,404	Del	B007VE2	Yes
chr22:22676385-22717669	6	41,285	Del	B007VDX	Yes
chr22:23999142-24244593	66	245,452	Del	B007VDC	Yes
chr22:22676385-22717669	6	41,285	Dup	B007VBT	Yes
chr22:22676385-22717669	6	41,285	Del	B007VB1	Yes
chr22:17257787-17388108	41	130,322	Del	B007V8F	Yes
chr22:24056992-24105834	14	48,843	Dup	B007V7Y	Yes
chr22:24159084-24239811	34	80,728	Dup	B007V7Y	Yes
chr22:17257787-17388108	41	130,322	Dup	B007V7K	Yes
chr22:22688572-22717669	5	29,098	Dup	B007V76	Yes
chr22:24128856-24244593	40	115,738	Dup	B007V76	Yes
chr22:22619365-22653131	5	33,767	Dup	B007V49	No
chr22:22664948-22698161	6	33,214	Del	B007V49	Yes
chr22:24016040-24240667	62	224,628	Dup	B007V45	Yes
chr22:21389563-21419339	11	29,777	Del	B007V42	Yes
chr22:22676385-22717669	6	41,285	Dup	B007V42	Yes
chr22:24037443-24240667	58	203,225	Dup	B007V42	Yes
chr22:22676385-22728586	7	52,202	Dup	B007V3P	Yes
chr22:22111075-22183571	24	72,497	Del	B007V35	Yes
chr22:20645312-20903637	190	258,326	Del	B007V1Y	Yes
chr22:23983992-24046526	12	62,535	Dup	B007V0B	Yes
chr22:24159084-24183368	8	24,285	Dup	B007V0B	Yes
chr22:24226979-24303664	31	76,686	Dup	B007V0B	Yes
chr22:21384614-21419339	12	34,726	Del	B007V04	Yes
chr22:23983992-24244593	69	260,602	Del	B007UYN	Yes
chr22:19066315-19366850	92	300,536	Del	B007UXY	Yes
chr22:19369581-19384193	6	14,613	Dup	B007UXY	Yes
chr22:19418146-19632925	27	214,780	Del	B007UXY	Yes
chr22:19710064-19792353	50	82,290	Del	B007UXY	Yes
chr22:17292678-17337129	15	44,452	Del	B007UXM	No
chr22:22059698-22095953	16	36,256	Del	B007UXM	No
chr22:24016040-24230823	55	214,784	Del	B007UXM	Yes
chr22:22668071-22717669	7	49,599	Del	B007UXB	Yes
chr22:21361579-21378145	6	16,567	Dup	B007UX8	No
chr22:17257787-17398986	43	141,200	Del	B007UX5	Yes
chr22:18104613-18118355	9	13,743	Del	B007UX5	No
chr22:18301003-18358558	32	57,556	Del	B007UX5	No
chr22:18101766-18185263	30	83,498	Dup	B007UWK	No
chr22:23999142-24046526	9	47,385	Del	B007UWC	Yes
chr22:24135224-24244593	39	109,370	Del	B007UWC	Yes
chr22:18145182-18185263	16	40,082	Del	B007UW1	No
chr22:18328337-18358558	22	30,222	Del	B007UW1	No
chr22:22676385-22717669	6	41,285	Dup	B007UVX	Yes
chr22:23994408-24244593	67	250,186	Del	B007UVQ	Yes
chr22:24037443-24238441	56	200,999	Dup	B007UVH	Yes
chr22:22676385-22717669	6	41,285	Dup	B007UVG	Yes
chr22:22676385-22717669	6	41,285	Dup	B007UUQ	Yes
chr22:17292678-17337129	15	44,452	Del	B007UTU	No

chr22:18101766-18115162	7	13,397	Del	B007UTU	No
chr22:22059698-22077079	6	17,382	Del	B007UTU	No
chr22:20590441-20604745	6	14,305	Del	B007UTL	No
chr22:17257787-17337129	20	79,343	Del	B007UT7	No
chr22:18328337-18358558	22	30,222	Del	B007UT7	No
chr22:18487729-18545632	13	57,904	Del	B007UT7	No
chr22:23983992-24244593	69	260,602	Dup	B007UT7	Yes
chr22:22619365-22653131	5	33,767	Del	B007USZ	No
chr22:18528368-18545632	7	17,265	Del	B007USH	No
chr22:17281004-17345628	20	64,625	Del	B007USB	No
chr22:22059698-22123801	21	64,104	Del	B007USB	No
chr22:22899468-22973609	14	74,142	Del	B007USB	No
chr22:21615252-21799918	33	184,667	Del	B007US1	No
chr22:23994408-24244593	67	250,186	Del	B007URR	Yes
chr22:18328337-18400229	39	71,893	Del	B007URL	No
chr22:22059698-22095953	16	36,256	Del	B007URL	No
chr22:23994408-24244593	67	250,186	Del	B007URJ	Yes
chr22:18101766-18118355	11	16,590	Del	B007UQP	No
chr22:21328337-21980987	140	652,651	Del	B007UQP	Yes
chr22:22676385-22717669	6	41,285	Dup	B007UQE	Yes
chr22:23994408-24244593	67	250,186	Del	B007UQ2	Yes
chr22:22619365-22653131	5	33,767	Dup	B007UPI	No
chr22:22664948-22698161	6	33,214	Del	B007UPI	Yes
chr22:22619365-22653131	5	33,767	Del	B007UPE	No
chr22:22676385-22717669	6	41,285	Dup	B007UPC	Yes
chr22:18780866-18821832	5	40,967	Dup	B007UOP	???
chr22:19094595-19114217	6	19,623	Del	B007UOD	No
chr22:19674884-19712482	26	37,599	Del	B007UOD	No
chr22:22619365-22698161	11	78,797	Del	B007UOD	Yes
chr22:22664948-22698161	6	33,214	Del	B007UO9	Yes
chr22:23994408-24240667	66	246,260	Dup	B007UNE	Yes
chr22:22676385-22717669	6	41,285	Dup	B007UMY	Yes
chr22:17257787-17388108	41	130,322	Del	B007UM2	Yes
chr22:17118296-18662886	378	1,544,591	Del	B007ULM	Yes
chr22:19066315-19791274	208	724,960	Del	B007ULM	Yes
chr22:18794653-18842490	5	47,838	Dup	B007UL1	???
chr22:21330030-21457585	36	127,556	Del	B007UKV	Yes
chr22:23994408-24136949	30	142,542	Dup	B007UKV	Yes
chr22:23999142-24238441	63	239,300	Dup	B007UK3	Yes
chr22:20645312-20903637	190	258,326	Dup	B007UJZ	Yes
chr22:23649537-23711045	12	61,509	Dup	B007UJZ	Yes
chr22:17241748-17388108	43	146,361	Dup	B007UJY	Yes
chr22:22688572-22717669	5	29,098	Dup	B007UJY	Yes
chr22:22676385-22717669	6	41,285	Del	B007UIU	Yes
chr22:20725035-20901854	157	176,820	Dup	B007UIS	Yes
chr22:22676385-22717669	6	41,285	Dup	B007UIS	Yes
chr22:22676385-22717669	6	41,285	Dup	B007UIR	Yes
chr22:23994408-24244593	67	250,186	Dup	B007UI6	Yes
chr22:22676385-22717669	6	41,285	Dup	B007UHM	Yes
chr22:22653131-22698161	7	45,031	Del	B007UH0	Yes
chr22:23994408-24240667	66	246,260	Del	B007UGU	Yes
chr22:22676385-22728586	7	52,202	Dup	B007UGR	Yes
chr22:22688572-22717669	5	29,098	Dup	B007UG7	Yes
chr22:22688572-22717669	5	29,098	Dup	B007UFX	Yes
chr22:23994408-24238441	64	244,034	Dup	B007UFS	Yes
chr22:17343340-17396663	22	53,324	Dup	B007UFP	Yes
chr22:23999142-24230823	58	231,682	Del	B007UEW	Yes
chr22:23991725-24178811	38	187,087	Dup	B007U9W	Yes
chr22:24217513-24238441	12	20,929	Dup	B007U9W	Yes
chr22:22676385-22717669	6	41,285	Del	B007U9T	Yes
chr22:22013796-22264030	85	250,235	Dup	B007U9J	Yes

chr22:22407117-22698161	84	291,045	Dup	B007U9J	Yes
chr22:22744829-23327473	100	582,645	Dup	B007U9J	Yes
chr22:23983992-24178811	39	194,820	Dup	B007U9J	Yes
chr22:24208614-24236803	14	28,190	Dup	B007U9J	Yes
chr22:22688572-22717669	5	29,098	Dup	B007U9C	Yes
chr22:24079793-24334249	75	254,457	Dup	B007U99	Yes
chr22:22676385-22717669	6	41,285	Dup	B007U90	Yes
chr22:22619365-22653131	5	33,767	Dup	B007U8X	No
chr22:22664948-22698161	6	33,214	Del	B007U8X	Yes
chr22:24198241-24228162	17	29,922	Dup	B007U6X	Yes

Table 8.8. Table of PennCNV called CNVs followed up in the 22q11.2 region with array CGH.

8.5 Chapter 5

8.5.3.1 CNV Calling

8.5.3.1.1 QuantiSNP

Perl script for automatic submission of QuantiSNP jobs, written by David To and

James Rucker

```
#!/usr/bin/perl
use strict;
use warnings;

my $fileList = 'filelist';
my $genderFile = “<gender_file>";
my %genderLookup;
open(GEN, $genderFile);
while(my $line = <GEN>)
{
	chomp($line);
	my($sampleId, $gender) = split(/,/, $line);
	if($gender eq "1")
	{
		$gender = 'male';
	}
	else
	{
		$gender = 'female';
	}
	$genderLookup{$sampleId} = $gender;
}
close(GEN);
```

```
open(LOG, ">genscripts.log");
open(LIST, $fileList);
while(my $line = <LIST>)
{
	chomp($line);
	my $outfile = "job-$line.sh";

	my $sampleId = $line;

	my $sex = 'unknown';
	if(defined($genderLookup{$sampleId}))
	{
		$sex = $genderLookup{$sampleId};
	}
	else
	{
		print LOG "Could not determine the sex for: $sampleId -
skipping\n";
		next;
	}
	open(OUT, ">$outfile");

	print OUT <<EOF;
#!/bin/sh
#\$-S /bin/sh
#\$-o /home/jrucker/sge-output
#\$-e /home/jrucker/sge-output
#\$-q short.q,long.q,bignode.q
#\$-l h_vmem=4G
#\$-N NBS_$sampleId
#\$-pe multi_thread 1
#\$-p -300
/share/apps/quantisnp_2.3/bin/run_quantisnp2.sh --sampleid $sampleId --
input-files /home/jrucker/scratch/incoming/nbs/$line --outdir
/home/jrucker/scratch/incoming/nbs/quantisnp_out --gender $sex --gcdir
/home/jrucker/b36 --levels /share/apps/quantisnp_2.3/config/levels.dat --
config /share/apps/quantisnp_2.3/config/params.dat --plot --genotype
EOF
	system("chmod 0755 $outfile");
	close(OUT);
	system("qsub $outfile");
}
close(LOG);
close(LIST);
```

Parameter	Value	Description
nComp	3	Number of mixture components
v	10	Degrees of freedom of Student t-distribution
nu_alpha	1	Shape parameter of Beta prior on outlier rate
nu_beta	1	Shape parameter of Beta prior on outlier rate
w_alpha	1.00E+04	Scale parameter of Dirichlet prior on genotype proportions
q_alpha	10	Scale parameter of Dirichlet prior on mixture proportions
tau	1.00E+04	Concentration parameter of Normal-Wishart prior
S_alpha	3	Scale parameter of Wishart prior on covariance matrix
S_alpha_homdel	3	Scale parameter of Wishart prior on covariance matrix (Homozygous Deletion)
longChromosome	2.00E+06	Characteristic length for normal state

Table 8.9. Params.dat File - QuantiSNP2

CopyNumber	MeanLogRRatio
0	-4
1	-0.55
2	0
2	0
3	0.39
4	0.53
5	4

Table 8.10. Levels.dat File - QuantiSNP2

8.5.4.2 Results from the Intersected Call Set

We present here the results of our analyses with the intersected call set derived from the intersection of the iPattern and QuantiSNP methods, with PennCNV used for confirmation. For brevity we present our standard QC and high QC analyses in this dataset side by side in each table.

8.5.4.2.1 Cases Vs. Screened Controls

We initially compared the results of each method in our cases and screened controls.

Table 8.11 illustrates the CNV event rate per sample. No significant differences are seen in this analysis.

Cases Vs. screened controls.		**Standard QC**			**High QC**		
Event Frequency	**Event Type**	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	0.67 | 0.74	0.93	0.15	0.64 | 0.68	0.81	0.39
Rare	**Del**	0.36 | 0.38	0.72	0.60	0.35 | 0.35	0.56	0.96
Rare	**Dup**	0.37 | 0.42	0.93	0.14	0.34 | 0.40	0.95	0.12

Table 8.11. CNV event rate per person (Rate). Significance values of less than 0.05 are highlighted in bold.

Table 8.12 illustrates the proportion of cases and controls to have at least 1 CNV event. There are no significant differences within this analysis.

Cases Vs. screened controls.		**Standard QC**			**High QC**		
Event Frequency	**Event Type**	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	0.48 | 0.50	0.80	0.46	0.47 | 0.48	0.59	0.86
Rare	**Del**	0.30 | 0.32	0.78	0.49	0.29 | 0.30	0.65	0.80
Rare	**Dup**	0.29 | 0.34	0.96	0.092	0.28 | 0.31	0.93	0.15

Table 8.12. Proportion of cases/controls to have at least one CNV event (Prop). Significance values of less than 0.05 are highlighted in bold.

Table 8.13 illustrates the total CNV event distance per subject. There are no significant differences within this analysis.

Cases Vs. screened controls.		Standard QC			High QC		
Event Frequency	**Event Type**	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	418 | 391	0.29	0.53	417 | 385	0.26	0.49
Rare	**Del**	306 | 260	0.15	0.31	313 | 265	0.18	0.35
Rare	**Dup**	425 | 379	0.17	0.32	421 | 389	0.26	0.51

Table 8.13. Total CNV event distance spanned per subject in kb (KbTot). Significance values of less than 0.05 are highlighted in bold.

Table 8.14 illustrates the average CNV event size per subject. There are no significant differences within this analysis.

Cases Vs. screened controls.		Standard QC			High QC		
Event Frequency	**Event Type**	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	290 | 272	0.23	0.44	293 | 276	0.27	0.50
Rare	**Del**	246 | 217	0.15	0.31	248 | 225	0.26	0.51
Rare	**Dup**	327 | 304	0.24	0.46	336 | 315	0.28	0.55

Table 8.14. Average CNV event size in kb per subject (KbAvg). Significance values of less than 0.05 are highlighted in bold.

Table 8.15 illustrates the number of genes spanned by CNVs. There are no significant differences in this analysis.

Cases Vs. screened controls.		Standard QC			High QC		
Event Frequency	**Event Type**	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	0.14 | 0.14	0.46	0.88	0.13 | 0.14	0.57	1.00
Rare	**Del**	0.073 | 0.067	0.38	0.75	0.077 | 0.066	0.31	0.52
Rare	**Dup**	0.071 | 0.072	0.59	1.00	0.072 | 0.075	0.63	0.91

Table 8.15. The number of genes spanned by CNV events (GRate). Significance values of less than 0.05 are highlighted in bold.

Table 8.16 illustrates the number of CNVs involving at least one gene There are no significant differences in this analysis.

Cases Vs. screened controls.		Standard QC			High QC		
Event Frequency	**Event Type**	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	0.13 | 0.13	0.51	1.00	0.13 | 0.13	0.58	1.00
Rare	**Del**	0.071 | 0.067	0.42	0.83	0.0750| 0.066	0.33	0.58
Rare	**Dup**	0.067 | 0.070	0.61	0.91	0.069 | 0.072	0.64	0.91

Table 8.16. The number of CNV events involving at least one gene (GProp). Significance values of less than 0.05 are highlighted in bold.

Finally, table 8.17 illustrates the number of genes within each kilobase of CNV.

There are no significant differencesin this analysis.

Cases Vs. screened controls.		Standard QC			High QC		
Event Frequency	**Event Type**	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	0.00123 | 0.000967	0.089	0.18	0.00123 | 0.00106	0.22	0.42
Rare	**Del**	0.00132 | 0.00104	0.15	0.30	0.00144 | 0.00111	0.13	0.26
Rare	**Dup**	0.00105 | 0.000843	0.20	0.40	0.00116 | 0.000940	0.20	0.40

Table 8.17. Number of genes per total CNV in kb (GRich). Significance values of less than 0.05 are highlighted in bold.

8.5.4.2.2 Cases Vs. WTCCC2 Controls

We then compared the results of our intersected call set in our high QC cases and the WTCCC2 controls.

Table 8.18 illustrates the CNV event rate per person. There are no significant differences in this analysis.

Cases Vs. WTCCC2 controls.		Standard QC			High QC		
Event Frequency	**Event Type**	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Rate (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	0.66 | 0.68	0.75	0.51	0.68 | 0.69	0.66	0.70
Rare	**Del**	0.34 | 0.34	0.50	1.00	0.37 | 0.35	0.15	0.29
Rare	**Dup**	0.39 | 0.39	0.70	0.61	0.38 | 0.40	0.87	0.27

Table 8.18. CNV event rate per person (Rate). Significance values of less than 0.05 are highlighted in bold.

Table 8.19 illustrates the proportion of subjects to have at least one CNV event.

There are no significant differences in this analysis.

Cases Vs. WTCCC2 controls.		Standard QC			High QC		
Event Frequency	**Event Type**	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**Prop (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	0.48 | 0.47	0.22	0.43	0.49 | 0.48	0.18	0.35
Rare	**Del**	0.29 | 0.29	0.49	0.96	0.30 | 0.29	0.23	0.43
Rare	**Dup**	0.30 | 0.30	0.54	0.95	0.30 | 0.31	0.69	0.65

Table 8.19. Proportion of cases/controls to have at least one CNV event (Prop). Significance values of less than 0.05 are highlighted in bold.

Table 8.20 illustrates the total CNV event distance in kilobases per subject. There are no significant differences in this analysis.

Cases Vs. WTCCC2 controls.		Standard QC			High QC		
Event Frequency	**Event Type**	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbTot (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	430 | 428	0.46	0.92	425 | 450	0.65	0.65
Rare	**Del**	308 | 282	0.078	0.15	320 | 292	0.10	0.19
Rare	**Dup**	436 | 442	0.61	0.79	424 | 472	0.85	0.41

Table 8.20. Total CNV event distance spanned per subject in kb (KbTot). Significance values of less than 0.05 are highlighted in bold.

Table 8.21 illustrates the average CNV event size in kilobases per subject. There are no significant differences in this analysis.

Cases Vs. WTCCC2 controls.		**Standard QC**			**High QC**		
Event Frequency	**Event Type**	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**KbAvg (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	299 | 293	0.30	0.59	293 | 295	0.55	0.89
Rare	**Del**	251 | 237	0.12	0.24	249 | 239	0.21	0.40
Rare	**Dup**	329 | 336	0.67	0.66	328 | 342	0.81	0.36

Table 8.21. Average CNV event size in kb per subject (KbAvg). Significance values of less than 0.05 are highlighted in bold.

Table 8.22 illustrates the number of genes spanned by CNV events in cases and WTCCC2 controls. There are no significant differences in this analysis.

Cases Vs. WTCCC2 controls.		**Standard QC**			**High QC**		
Event Frequency	**Event Type**	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRate (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	0.14 | 0.51	0.82	0.40	0.14 | 0.15	0.68	0.67
Rare	**Del**	0.073 | 0.079	0.82	0.40	0.077 | 0.079	0.63	0.79
Rare	**Dup**	0.071 | 0.074	0.69	0.66	0.072 | 0.075	0.68	0.68

Table 8.22. The number of genes spanned by CNV events (GRate). Significance values of less than 0.05 are highlighted in bold.

Table 8.23 illustrates the number of CNV events to involve at least one gene. There are no significant differences in this analysis.

Cases Vs. WTCCC2 controls.		**Standard QC**			**High QC**		
Event Frequency	**Event Type**	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GProp (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	0.13 | 0.14	0.67	0.70	0.14 | 0.14	0.51	1.00
Rare	**Del**	0.071 | 0.076	0.76	0.52	0.075 | 0.076	0.56	0.93

Rare	**Dup**	0.068 | 0.068	0.56	0.92	0.069 | 0.070	0.54	0.96

Table 8.23. The number of CNV events involving at least one gene (GProp). Significance values of less than 0.05 are highlighted in bold.

Finally, table 8.24 illustrates the number of genes per kilobase of CNV in cases and controls. There are no significant differences in this analysis.

Cases Vs. WTCCC2 controls.		**Standard QC**			**High QC**		
Event Frequency	**Event Type**	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**	**GRich (cases | controls)**	**1 sided p val.**	**2 sided p val.**
Rare	**Del/Dup**	0.00125 | 0.00122	0.38	0.75	0.00127 | 0.00121	0.27	0.52
Rare	**Del**	0.00139 | 0.00140	0.52	0.94	0.00137 | 0.00136	0.43	0.89
Rare	**Dup**	0.00100 | 0.000944	0.28	0.55	0.00102 | 0.000954	0.25	0.51

Table 8.24. Number of genes per total CNV in kb (GRich). Significance values of less than 0.05 are highlighted in bold.

References

American Psychiatric Association. (1994). *DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Press Inc.

Andreasen, N. (1995). Symptoms, signs, and diagnosis of schizophrenia. *The Lancet*, *346*(8973), 477–481. doi:10.1016/S0140-6736(95)91325-4

Babatz, T. D., Kumar, R. A., Sudi, J., Dobyns, W. B., & Christian, S. L. (2009). Copy number and sequence variants implicate APBA2 as an autism candidate gene. *Autism Research: Official Journal of the International Society for Autism Research*, *2*(6), 359–364. doi:10.1002/aur.107

Bailey, A., Le Couteur, A., Gottesman, I., Bolton, P., Simonoff, E., Yuzda, E., & Rutter, M. (1995). Autism as a strongly genetic disorder: evidence from a British twin study. *Psychological Medicine*, *25*(1), 63–77.

Bailey, A., Phillips, W., & Rutter, M. (1996). Autism: towards an integration of clinical, genetic, neuropsychological, and neurobiological perspectives. *Journal of Child Psychology and Psychiatry*, *37*(1), 89–126.

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., & Eichler, E. E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome Research*, *11*(6), 1005–1017. doi:10.1101/gr.187101

Balciuniene, J., Feng, N., Iyadurai, K., Hirsch, B., Charnas, L., Bill, B. R., Easterday, M. C., et al. (2007). Recurrent 10q22-q23 Deletions: A Genomic Disorder on 10q Associated with Cognitive and Behavioral Abnormalities. *American*

Journal of Medical Genetics, *80*(5), 938–947. doi:10.1086/513607

Ballif, B. C., Theisen, A., Coppinger, J., Gowans, G. C., Hersh, J. H., Madan-Khetarpal, S., Schmidt, K. R., et al. (2008). Expanding the clinical phenotype of the 3q29 microdeletion syndrome and characterization of the reciprocal microduplication. *Molecular Cytogenetics*, *1*, 8. doi:10.1186/1755-8166-1-8

Baranzini, S. E., Wang, J., Gibson, R. A., Galwey, N., Naegelin, Y., Barkhof, F., Radue, E.-W., et al. (2009). Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Human Molecular Genetics*, *18*(4), 767–778. doi:10.1093/hmg/ddn388

Bender, B., Harmon, R., Linden, M., & Robinson, A. (1995). Psychosocial adaptation of 39 adolescents with sex chromosome abnormalities. *Pediatrics*, *96*(2 Pt 1), 302–308.

Benjamin, J., Ebstein, R. P., & Belmaker, R. H. (2002). *Molecular genetics and the human personality* (p. 356). American Psychiatric Pub.

Berg, J. S., Brunetti-Pierri, N., Peters, S. U., Kang, S.-H. L., Fong, C.-T., Salamone, J., Freedenberg, D., et al. (2007). Speech delay and autism spectrum behaviors are frequently associated with duplication of the 7q11.23 Williams-Beuren syndrome region. *Genetics in Medicine*, *9*(7), 427–441. doi:10.1097/GIM.0b013e3180986192

Biederer, T., & Südhof, T. C. (2000). Mints as adaptors. Direct binding to neurexins and recruitment of munc18. *Journal of Biological Chemistry*, *275*(51), 39803–39806. doi:10.1074/jbc.C000656200

Biederman, J. (1998). Attention-deficit/hyperactivity disorder: a life-span perspective. *Journal of Clinical Psychiatry*, *59*(S7), 4–16.

Bierut, L. J., Heath, A. C., Bucholz, K. K., Dinwiddie, S. H., Madden, P. A., Statham, D. J., Dunne, M. P., et al. (1999). Major depressive disorder in a community-based twin sample: are there different genetic and environmental contributions for men and women? *Archives of General Psychiatry*, *56*(6), 557–563.

Binder, E. B., Kinkead, B., Owens, M. J., & Nemeroff, C. B. (2001). Neurotensin and dopamine interactions. *Pharmacological Reviews*, *53*(4), 453–486.

Bochukova, E. G., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K., Saeed, S., et al. (2009). Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, *463*(7281), 666–670. doi:10.1038/nature08689

Breen, G., Webb, B. T., Butler, A. W., van den Oord, E. J. C. G., Tozzi, F., Craddock, N., Gill, M., et al. (2011). A Genome-Wide Significant Linkage for Severe Depression on Chromosome 3: The Depression Network Study. *American Journal of Psychiatry*, *168*(8), 840–847. doi:10.1176/appi.ajp.2011.10091342

Brzustowicz, L. M., Hodgkinson, K. A., Chow, E. W., Honer, W. G., & Bassett, A. S. (2000). Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-q22. *Science*, *288*(5466), 678–682.

Buizer-Voskamp, J. E., Muntjewerff, J.-W., Genetic Risk and Outcome in Psychosis (GROUP) Consortium Members, Strengman, E., Sabatti, C., Stefansson, H., Vorstman, J. A. S., et al. (2011). Genome-wide analysis shows increased frequency of copy number variation deletions in Dutch schizophrenia patients. *Biological Psychiatry*, *70*(7), 655–662. doi:10.1016/j.biopsych.2011.02.015

Burn, J., & Goodship, J. (1996). Developmental genetics of the heart. *Curr Opin Genet Dev*, *6*(3), 322–325.

Butler, A., Cohen-Woods, S., Farmer, A., McGuffin, P., & Lewis, C. (2010). Integrating phenotypic data for depression. *Journal of Integrative Bioinformatics*, *7*(3). doi:10.2390/biecoll-jib-2010-136

Cadoret, R., O'Gorman, T., Heywood, E., & Troughton, E. (1985). Genetic and environmental factors in major depression. *Journal of Affective Disorders*, *9*(2), 155–164.

Camp, N., Lowry, M., Richards, R., Plenk, A., Carter, C., Hensel, C., Abkevich, V., et al. (2005). Genome-wide linkage analyses of extended Utah pedigrees identifies loci that influence recurrent, early-onset major depression and anxiety disorders. *American Journal of Medical Genetics: Part B Neuropsychiatric Genetics*, *135B*(1), 85–93. doi:10.1002/ajmg.b.30177

Cichon, S., Craddock, N., Daly, M., Faraone, S., Gejman, P., Kelsoe, J., Lehner, T., et al. (2009). Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *American Journal of Psychiatry*, *166*(5), 540–556. doi:10.1176/appi.ajp.2008.08091354

Cichon, S., Mühleisen, T. W., Degenhardt, F. A., Mattheisen, M., Miró, X., Strohmaier, J., Steffens, M., et al. (2011). Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. *American Journal of Medical Genetics*, *88*(3), 372–381. doi:10.1016/j.ajhg.2011.01.017

Clayton-Smith, J., Giblin, C., Smith, R. A., Dunn, C., & Willatt, L. (2010). Familial 3q29 microdeletion syndrome providing further evidence of involvement of

the 3q29 region in bipolar disorder. *Clinical Dysmorphology*, *19*(3), 128–132. doi:10.1097/MCD.0b013e32833a1e3c

Colella, S., Yau, C., Taylor, J., Mirza, G., Butler, H., Clouston, P., Bassett, A., et al. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*, *35*(6), 2013–2025. doi:10.1093/nar/gkm076

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, *464*(7289), 704–712. doi:10.1038/nature08516

Craddock, N., & Forty, L. (2006). Genetics of affective (mood) disorders. *European Journal of Human Genetics*, *14*(6), 660–668. doi:10.1038/sj.ejhg.5201549

Craddock, N., & Jones, I. (1999). Genetics of bipolar disorder. *Journal of Medical Genetics*, *36*(8), 585–594.

Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., Vukcevic, D., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, *464*(7289), 713–720. doi:10.1038/nature08979

Cuscó, I., Medrano, A., Gener, B., Vilardell, M., Gallastegui, F., Villa, O., González, E., et al. (2009). Autism-specific copy number variants further implicate the phosphatidylinositol signaling pathway and the glutamatergic synapse in the etiology of the disorder. *Human Molecular Genetics*, *18*(10), 1795–1804. doi:10.1093/hmg/ddp092

de Vries, B. B. A., Pfundt, R., Leisink, M., Koolen, D. A., Vissers, L. E. L. M., Janssen, I.

M., Reijmersdal, S. V., et al. (2005). Diagnostic genome profiling in mental retardation. *American Journal of Medical Genetics*, *77*(4), 606–616. doi:10.1086/491719

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, *3*(2), 71–100. doi:10.1177/1534582304267187

Degenhardt, F., Priebe, L., Herms, S., Mattheisen, M., Mühleisen, T. W., Meier, S., Moebus, S., et al. (2012). Association between copy number variants in 16p11.2 and major depressive disorder in a German case-control sample. *American Journal of Medical Genetics: Part B Neuropsychiatric Genetics*. doi:10.1002/ajmg.b.32034

Dellinger, A. E., Saw, S.-M., Goh, L. K., Seielstad, M., Young, T. L., & Li, Y.-J. (2010). Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Research*, *38*(9), e105. doi:10.1093/nar/gkq040

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977, March 28). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methdological).* Retrieved March 12, 2012, from http://www.jstor.org/discover/10.2307/2984875?uid=3738032&uid=2&uid=4&sid=47698747032667

Diskin, S., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., et al. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research*, *36*(19), e126. doi:10.1093/nar/gkn556

Dulubova, I., Khvotchev, M., Liu, S., Huryeva, I., Südhof, T. C., & Rizo, J. (2007). Munc18-1 binds directly to the neuronal SNARE complex. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(8), 2697–2702. doi:10.1073/pnas.0611318104

Dunham, I., Hunt, A. R., Collins, J. E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L. J., et al. (1999). The DNA sequence of human chromosome 22. *Nature*, *402*(6761), 489–495. doi:10.1038/990031

Eckel-Passow, J. E., Atkinson, E. J., Maharjan, S., Kardia, S. L., & de Andrade, M. (2011). Software comparison for evaluating genomic copy number variation for Affymetrix 6.0 SNP array platform. *BMC Bioinformatics*, *12*(1), 220. doi:10.1186/1471-2105-12-220

Eichler, E. E. (2001). Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends in Genetics*, *17*(11), 661–669.

Elia, J., Gai, X., Xie, H. M., Perin, J. C., Geiger, E., Glessner, J. T., D'arcy, M., et al. (2010). Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Molecular Psychiatry*, *15*(6), 637–646. doi:10.1038/mp.2009.57

Elsea, S. H., & Girirajan, S. (2008). Smith–Magenis syndrome. *European Journal of Human Genetics*, *16*(4), 412–421. doi:10.1038/sj.ejhg.5202009

Emanuel, B. S. (2008). Molecular Mechanisms and Diagnosis of Chromosome 22Q11.2 Rearrangements. *Developmental Disabilities Research Reviews*, *14*(1), 11. doi:10.1002/ddrr.3

Engels, H., Brockschmidt, A., Hoischen, A., Landwehr, C., Bosse, K., Walldorf, C., Toedt, G., et al. (2007). DNA microarray analysis identifies candidate regions

and genes in unexplained mental retardation. *Neurology*, *68*(10), 743–750. doi:10.1212/01.wnl.0000256367.70365.e0

Eysenck, H., & Eysenck, S. (1964). *Manual of the Eysenck Personality Questionnaire* (1st ed.). London: Hodder and Stoughton.

Fallin, M. D., Lasseter, V. K., Wolyniec, P. S., McGrath, J. A., Nestadt, G., Valle, D., Liang, K. Y., et al. (2003). Genomewide linkage scan for schizophrenia susceptibility loci among Ashkenazi Jewish families shows evidence of linkage on chromosome 10q22. *American Journal of Medical Genetics*, *73*(3), 601–611.

Fan, Y.-S., Jayakar, P., Zhu, H., Barbouth, D., Sacharow, S., Morales, A., Carver, V., et al. (2007). Detection of pathogenic gene copy number variations in patients with mental retardation by genomewide oligonucleotide array comparative genomic hybridization. *Hum Mutat*, *28*(11), 1124–1132. doi:10.1002/humu.20581

Faraone, S. V., Hwu, H. G., Liu, C. M., Chen, W. J., Tsuang, M. M., Liu, S. K., Shieh, M. H., et al. (2006). Genome Scan of Han Chinese Schizophrenia Families From Taiwan: Confirmation of Linkage to 10q22.3. *American Journal of Psychiatry*, *163*(10), 1760–1766. doi:10.1176/appi.ajp.163.10.1760

Farmer, A., Breen, G., Brewster, S., Craddock, N., Gill, M., Korszun, A., Maier, W., et al. (2004). The Depression Network (DeNT) Study: methodology and sociodemographic characteristics of the first 470 affected sibling pairs from a large multi-site linkage genetic study. *BMC Psychiatry*, *4*, 42. doi:10.1186/1471-244X-4-42

Farmer, A., Harris, T., Redman, K., Sadler, S., Mahmood, A., & McGuffin, P. (2000).

Cardiff depression study. A sib-pair study of life events and familiality in major depression. *British Journal of Psychiatry*, *176*, 150–155.

Fava, M., Uebelacker, L. A., Alpert, J. E., Nierenberg, A. A., Pava, J. A., & Rosenbaum, J. F. (1997). Major depressive subtypes and treatment response. *Biological Psychiatry*, *42*(7), 568–576. doi:10.1016/S0006-3223(96)00440-4

Feng, J., Schroer, R., Yan, J., Song, W., Yang, C., Bockholt, A., Cookjr, E., et al. (2006). High frequency of neurexin 1β signal peptide structural variants in patients with autism. *Neuroscience Letters*, *409*(1), 10–13. doi:10.1016/j.neulet.2006.08.017

Fernandez, B. A., Roberts, W., Chung, B., Weksberg, R., Meyn, S., Szatmari, P., Joseph-George, A. M., et al. (2010). Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. *Journal of Medical Genetics*, *47*(3), 195–203. doi:10.1136/jmg.2009.069369

Ferreira, M., O'Donovan, M., Meng, Y., Jones, I., Ruderfer, D., Jones, L., Fan, J., et al. (2008). Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nature Genetics*, *40*(9), 1056–1058. doi:10.1038/ng.209

Feuk, L., Carson, A., & Scherer, S. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, *7*(2), 85–97. doi:10.1038/nrg1767

Fisher, R. (n.d.). On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society. Series B (Methodological).* Retrieved March 28, 1922, from http://www.jstor.org/stable/2340521?origin=crossref

Franke, B., Neale, B. M., & Faraone, S. V. (2009). Genome-wide association studies in ADHD. *Human Genetics*, *126*(1), 13–50. doi:10.1007/s00439-009-0663-4

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., et al. (2006). Copy number variation: new insights in genome diversity. *Genome Research*, *16*(8), 949–961. doi:10.1101/gr.3677206

Freitag, C. M. (2007). The genetics of autistic disorders and its clinical relevance: a review of the literature. *Molecular Psychiatry*, *12*(1), 2–22. doi:10.1038/sj.mp.4001896

Friedman, J. M., Baross, A., Delaney, A. D., Ally, A., Arbour, L., Asano, J., Bailey, D. K., et al. (2006). Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation. *American Journal of Medical Genetics*, *79*(3), 500–513. doi:10.1086/507471

Frodl, T., Meisenzahl, E. M., Zetzsche, T., Born, C., Groll, C., Jäger, M., Leinsinger, G., et al. (2002). Hippocampal changes in patients with a first episode of major depression. *American Journal of Psychiatry*, *159*(7), 1112–1118.

Gaysina, D., Cohen, S., Craddock, N., Farmer, A., Hoda, F., Korszun, A., Owen, M., et al. (2008). No association with the 5,10-methylenetetrahydrofolate reductase gene and major depressive disorder: results of the depression case control (DeCC) study and a meta-analysis. *American Journal of Medical Genetics: Part B Neuropsychiatric Genetics*, *147B*(6), 699–706. doi:10.1002/ajmg.b.30665

Gaysina, D., Cohen-Woods, S., Chow, P., Martucci, L., Schosser, A., Ball, H., Tozzi, F., et al. (2009). Association of the dystrobrevin binding protein 1 gene (DTNBP1) in a bipolar case-control study (BACCS). *American Journal of*

Medical Genetics: Part B Neuropsychiatric Genetics, *150B*(6), 836–844. doi:10.1002/ajmg.b.30906

Genetic SNP information from the 1958 British Birth Cohort. (n.d.). Genetic SNP information from the 1958 British Birth Cohort. *b58cgene.sgul.ac.uk*. Retrieved December 5, 2011, from http://www.b58cgene.sgul.ac.uk/acknowledgment.php

Gershon, E. S., Hamovit, J., Guroff, J. J., Dibble, E., Leckman, J. F., Sceery, W., Targum, S. D., et al. (1982). A Family Study of Schizoaffective, Bipolar I, Bipolar II, Unipolar, and Normal Control Probands. *Archives of General Psychiatry*, *39*(10), 1157. doi:10.1001/archpsyc.1982.04290100031006

Glessner, J., Wang, K., Sleiman, P., Zhang, H., Kim, C., Flory, J., Bradfield, J., et al. (2010). Duplication of the SLIT3 locus on 5q35.1 predisposes to major depressive disorder. *PLoS One*, *5*(12), e15463. doi:10.1371/journal.pone.0015463

Goodship, J., Cross, I., LiLing, J., & Wren, C. (1998). A population study of chromosome 22q11 deletions in infancy. *Archives of Diseases in Childhood*, *79*(4), 348–351.

Gottesman, I. I. (1991). *Schizophrenia genesis: The origins of madness.* W H Freeman/Times Books/ Henry Holt & Co.

Green, E. K., Grozeva, D., Jones, I., Jones, L., Kirov, G., Caesar, S., Gordon-Smith, K., et al. (2009). The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. *Molecular Psychiatry*, *15*(10), 1016–1022. doi:10.1038/mp.2009.49

Grozeva, D., Conrad, D. F., Barnes, C. P., Hurles, M., Owen, M. J., O'Donovan, M. C.,

Craddock, N., et al. (2012). Independent estimation of the frequency of rare CNVs in the UK population confirms their role in schizophrenia. *Schizophrenia Research*, *135*(1-3), 1–7. doi:10.1016/j.schres.2011.11.004

Grozeva, D., Kirov, G., Ivanov, D., Jones, I., Jones, L., Green, E., St Clair, D., et al. (2010). Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Archives of General Psychiatry*, *67*(4), 318–327. doi:10.1001/archgenpsychiatry.2010.25

Guilmatre, A., Dubourg, C., Mosca, A. L., Legallic, S., Goldenberg, A., Drouin-Garraud, V., Layet, V., et al. (2009). Recurrent Rearrangements in Synaptic and Neurodevelopmental Genes and Shared Biologic Pathways in Schizophrenia, Autism, and Mental Retardation. *Archives of General Psychiatry*, *66*(9), 947–956. doi:10.1001/archgenpsychiatry.2009.80

Gunderson, K., Kruglyak, S., Graige, M., Garcia, F., Kermani, B., Zhao, C., Che, D., et al. (2004). Decoding randomly ordered DNA arrays. *Genome Research*, *14*(5), 870–877. doi:10.1101/gr.2255804

Gunderson, K., Steemers, F., Lee, G., Mendoza, L., & Chee, M. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genetics*, *37*(5), 549–554. doi:10.1038/ng1547

Guo, S., Shi, Y., Zhao, X., Duan, S., Zhou, J., Meng, J., Yang, Y., et al. (2004). No genetic association between polymorphisms in the AMPA receptor subunit GluR4 gene (GRIA4) and schizophrenia in the Chinese population. *Neuroscience Letters*, *369*(2), 168–172. doi:10.1016/j.neulet.2004.07.079

Hahn, M. K., Blackford, J. U., Haman, K., Mazei-Robison, M., English, B. A., Prasad, H. C., Steele, A., et al. (2008). Multivariate permutation analysis associates

multiple polymorphisms with subphenotypes of major depression. *Genes, Brain and Behaviour*, *7*(4), 487–495. doi:10.1111/j.1601-183X.2007.00384.x

Harris, E., & Barraclough, B. (1998). Excess mortality of mental disorder. *British Journal of Psychiatry*, *173*, 11–53.

Hashimoto, R., Okada, T., Kato, T., Kosuga, A., Tatsumi, M., Kamijima, K., & Kunugi, H. (2005). The breakpoint cluster region gene on chromosome 22q11 is associated with bipolar disorder. *Biological Psychiatry*, *57*(10), 1097–1102. doi:10.1016/j.biopsych.2005.02.019

Hodgkinson, C. A., Goldman, D., Jaeger, J., Persaud, S., Kane, J. M., Lipsky, R. H., & Malhotra, A. K. (2004). Disrupted in schizophrenia 1 (DISC1): association with schizophrenia, schizoaffective disorder, and bipolar disorder. *American Journal of Medical Genetics*, *75*(5), 862–872. doi:10.1086/425586

Holmans, P., Zubenko, G., Crowe, R., Depaulo, J. J., Scheftner, W., Weissman, M., Zubenko, W., et al. (2004). Genomewide significant linkage to recurrent, early-onset major depressive disorder on chromosome 15q. *American Journal of Medical Genetics*, *74*(6), 1154–1167. doi:10.1086/421333

Iafrate, A., Feuk, L., Rivera, M., Listewnik, M., Donahoe, P., Qi, Y., Scherer, S., et al. (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, *36*(9), 949–951. doi:10.1038/ng1416

Inoue, K., & Lupski, J. R. (2003). Molecular Mechanisms for Genomic Disorders. *Annual Review of Genomics and Human Genetics*, *3*(1), 199–242. doi:10.1146/annurev.genom.3.032802.120023

International Schizophrenia Consortium. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, *455*(7210), 237–241.

doi:10.1038/nature07239

Jaillard, S., Drunat, S., Bendavid, C., Aboura, A., Etcheverry, A., Journel, H., Delahaye, A., et al. (2010). Identification of gene copy number variations in patients with mental retardation using array-CGH: Novel syndromes in a large French series. *European Journal of Human Genetics*, *53*(2), 66–75. doi:10.1016/j.ejmg.2009.10.002

Janowsky, D. S., El-Yousef, M. K., & Davis, J. M. (1974). Acetylcholine and depression. *Psychosomatic Medicine*, *36*(3), 248–257.

Jolin, E. M., Weller, R. A., Jessani, N. R., Zackai, E. H., McDonald-McGinn, D. M., & Weller, E. B. (2009). Affective disorders and other psychiatric diagnoses in children and adolescents with 22q11.2 Deletion Syndrome. *Journal of Affective Disorders*, *119*(1-3), 177–180. doi:10.1016/j.jad.2009.02.016

Jones, I., Kent, L., & Craddock, N. (2002). *Genetics of Affective Disorders*. Oxford Univeristy Press.

Judd, L. L., & Schettler, P. J. (2010). *Bipolar Disorder: Clinical and Neurobiological Foundations*. (L. N. Yatham & M. Maj, Eds.)The Long Term Course and Clinical Management of Bipolar I and Bipolar II Disorders (1st ed. p. 522). Wiley.

Kallioniemi, A., Kallioniemi, O., Sudar, D., Rutovitz, D., Gray, J., Waldman, F., & Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, *258*(5083), 818–821. doi:10.1126/science.1359641

Karayiorgou, M., Simon, T. J., & Gogos, J. A. (2010). 22q11.2 microdeletions: linking DNA structural variation to brain dysfunction and schizophrenia. *Nature Reviews Neuroscience*, *11*(6), 402–416. doi:10.1038/nrn2841

Keller, M. B., Shapiro, R. W., Lavori, P. W., & Wolfe, N. (1982). Relapse in major depressive disorder: analysis with the life table. *Archives of General Psychiatry*, *39*(8), 911–915.

Kendler, K. S., & Prescott, C. A. (1999). A population-based twin study of lifetime major depression in men and women. *Archives of General Psychiatry*, *56*(1), 39–44.

Kendler, K. S., Bulik, C. M., Silberg, J., Hettema, J. M., Myers, J., & Prescott, C. A. (2000). Childhood sexual abuse and adult psychiatric and substance use disorders in women: an epidemiological and cotwin control analysis. *Archives of General Psychiatry*, *57*(10), 953–959.

Kendler, K. S., Gardner, C. O., & Prescott, C. A. (1999). Clinical characteristics of major depression that predict risk of depression in relatives. *Archives of General Psychiatry*, *56*(4), 322–327.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C., & Eaves, L. J. (1992). A population-based twin study of major depression in women. The impact of varying definitions of illness. *Archives of General Psychiatry*, *49*(4), 257–266.

Kendler, K. S., Pedersen, N. L., Neale, M. C., & Mathé, A. A. (1995). A pilot Swedish twin study of affective illness including hospital- and population-ascertained subsamples: results of model fitting. *Behavioural Genetics*, *25*(3), 217–232.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, A. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. doi:10.1101/gr.229102

Kessing, L. V., Hansen, M. G., Andersen, P. K., & Angst, J. (2004). The predictive effect of episodes on the risk of recurrence in depressive and bipolar

disorders - a life-long perspective. *Acta Psychiatrica Scandinavica*, *109*(5), 339–344. doi:10.1046/j.1600-0447.2003.00266.x

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H.-U., et al. (1994). Lifetime and 12-Month Prevalence of DSM-III-R Psychiatric Disorders in the United States: Results From the National Comorbidity Survey. *Archives of General Psychiatry*, *51*(1), 8. doi:10.1001/archpsyc.1994.03950010008002

Kim, H.-G., Kishikawa, S., Higgins, A. W., Seong, I.-S., Donovan, D. J., Shen, Y., Lally, E., et al. (2008). Disruption of neurexin 1 associated with autism spectrum disorder. *American Journal of Medical Genetics*, *82*(1), 199–207. doi:10.1016/j.ajhg.2007.09.011

Kirov, G. (2010). The role of copy number variation in schizophrenia. *Expert Review of Neurotherapeutics*, *10*(1), 25–32. doi:10.1586/ern.09.133

Kirov, G., Grozeva, D., Norton, N., Ivanov, D., Mantripragada, K., Holmans, P., Craddock, N., Owen, M., et al. (2009a). Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Human Molecular Genetics*, *18*(8), 1497–1503. doi:10.1093/hmg/ddp043

Kirov, G., Gumus, D., Chen, W., Norton, N., Georgieva, L., Sari, M., O'Donovan, M. C., et al. (2007). Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Human Molecular Genetics*, *17*(3), 458–465. doi:10.1093/hmg/ddm323

Kirov, G., Rujescu, D., Ingason, A., Collier, D. A., O'Donovan, M. C., & Owen, M. J. (2009b). Neurexin 1 (NRXN1) Deletions in Schizophrenia. *Schizophrenia Bulletin*, *35*(5), 851–854. doi:10.1093/schbul/sbp079

Klauck, S. M. (2006). Genetics of autism spectrum disorder. *European Journal of Human Genetics*, *14*(6), 714–720. doi:10.1038/sj.ejhg.5201610

Knorring, von, A.-L., Cloninger, C. R., Bohman, M., & Sigvardsson, S. (1983). An Adoption Study of Depressive Disorders and Substance Abuse. *Archives of General Psychiatry*, *40*(9), 943. doi:10.1001/archpsyc.1983.01790080025003

Kohli, M. A., Lucae, S., Saemann, P. G., Schmidt, M. V., Demirkan, A., Hek, K., Czamara, D., et al. (2011). The Neuronal Transporter Gene SLC6A15 Confers Risk to Major Depression. *Neuron*, *70*(2), 252–265. doi:10.1016/j.neuron.2011.04.005

Korn, J., Kuruvilla, F., McCarroll, S., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., et al. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics*, *40*(10), 1253–1260. doi:10.1038/ng.237

Korszun, A., Moskvina, V., Brewster, S., Craddock, N., Ferrero, F., Gill, M., Jones, I. R., et al. (2004). Familiality of Symptom Dimensions in Depression. *Archives of General Psychiatry*, *61*(5), 468. doi:10.1001/archpsyc.61.5.468

La Rochebrochard, de, C., Joly-Hélas, G., Goldenberg, A., Durand, I., Laquerrière, A., Ickowicz, V., Saugier-Veber, P., et al. (2006). The intrafamilial variability of the 22q11.2 microduplication encompasses a spectrum from minor cognitive deficits to severe congenital anomalies. *American Journal of Medical Genetics*, *140A*(14), 1608–1613. doi:10.1002/ajmg.a.31227

Lander, E. S., & Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, *265*(5181), 2037–2048.

Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. doi:10.1038/35057062

Langley, K., Martin, J., Agha, S. S., Davies, C., Stergiakouli, E., Holmans, P., Williams, N., et al. (2011). Clinical and cognitive characteristics of children with attention-deficit hyperactivity disorder, with and without copy number variants. *British Journal of Psychiatry*, *199*, 398–403. doi:10.1192/bjp.bp.111.092130

Lesch, K.-P., Selch, S., Renner, T. J., Jacob, C., Nguyen, T. T., Hahn, T., Romanos, M., et al. (2011). Genome-wide copy number variation analysis in attention-deficit/hyperactivity disorder: association with neuropeptide Y gene dosage in an extended pedigree. *Molecular Psychiatry*, *16*(5), 491–503. doi:10.1038/mp.2010.29

Levinson, D. F., Duan, J., Oh, S., Wang, K., Sanders, A. R., Shi, J., Zhang, N., et al. (2011). Copy Number Variants in Schizophrenia: Confirmation of Five Previous Findings and New Evidence for 3q29 Microdeletions and VIPR2 Duplications. *American Journal of Psychiatry*, *168*(3), 302–316. doi:10.1176/appi.ajp.2010.10060876

Lewis, C., Ng, M., Butler, A., Cohen-Woods, S., Uher, R., Pirlo, K., Weale, M., et al. (2010). Genome-wide association study of major recurrent depression in the U.K. population. *American Journal of Psychiatry*, *167*(8), 949–957. doi:10.1176/appi.ajp.2010.09091380

Lieber, M. R., Yu, K., & Raghavan, S. C. (2006). Roles of nonhomologous DNA end joining, V(D)J recombination, and class switch recombination in

chromosomal translocations. *DNA repair*, *5*(9-10), 1234–1245. doi:10.1016/j.dnarep.2006.05.013

Lobo, I. (2008, March). Same Genetic Mutation, Different Genetic Disease Phenotype. *Nature Education*. Retrieved March 13, 2012, from

Locke, D., Segraves, R., Carbone, L., Archidiacono, N., Albertson, D., Pinkel, D., & Eichler, E. (2003). Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Research*, *13*(3), 347–357. doi:10.1101/gr.1003303

Longman, C. (2003). Mutations in the human LARGE gene cause MDC1D, a novel form of congenital muscular dystrophy with severe mental retardation and abnormal glycosylation of alpha-dystroglycan. *Human Molecular Genetics*, *12*(21), 2853–2861. doi:10.1093/hmg/ddg307

Lupski, J. (2007). An evolution revolution provides further revelation. *Bioessays*, *29*(12), 1182–1184. doi:10.1002/bies.20686

Lyons, M. J., Eisen, S. A., Goldberg, J., True, W., Lin, N., Meyer, J. M., Toomey, R., et al. (1998). A registry-based twin study of depression in men. *Archives of General Psychiatry*, *55*(5), 468–472.

Maier, W., Lichtermann, D., Minges, J., Hallmayer, J., Heun, R., Benkert, O., & Levinson, D. F. (1993). Continuity and discontinuity of affective disorders and schizophrenia. Results of a controlled family study. *Archives of General Psychiatry*, *50*(11), 871–883.

Makino, C., Fujii, Y., Kikuta, R., Hirata, N., Tani, A., Shibata, A., Ninomiya, H., et al. (2003). Positive association of the AMPA receptor subunit GluR4 gene (GRIA4) haplotype with schizophrenia: linkage disequilibrium mapping

using SNPs evenly distributed across the gene region. *American Journal of Medical Genetics: Part B Neuropsychiatric Genetics*, *116B*(1), 17–22. doi:10.1002/ajmg.b.10041

Malhotra, D., McCarthy, S., Michaelson, J. J., Vacic, V., Burdick, K. E., Yoon, S., Cichon, S., et al. (2011). High Frequencies of De Novo CNVs in Bipolar Disorder and Schizophrenia. *Neuron*, *72*(6), 951–963. doi:10.1016/j.neuron.2011.11.007

Marioni, J., Thorne, N., & Tavare, S. (2006). BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, *22*(9), 1144–1146. doi:10.1093/bioinformatics/btl089

Marshall, C., Noor, A., Vincent, J., Lionel, A., Feuk, L., Skaug, J., Shago, M., et al. (2008). Structural variation of chromosomes in autism spectrum disorder. *American Journal of Medical Genetics*, *82*(2), 477–488. doi:10.1016/j.ajhg.2007.12.009

Martens, M. A., Wilson, S. J., & Reutens, D. C. (2008). Research Review: Williams syndrome: a critical review of the cognitive, behavioral, and neuroanatomical phenotype. *Journal of Child Psychology and Psychiatry*, *49*(6), 576–608. doi:10.1111/j.1469-7610.2008.01887.x

Mathers, C., Fat, D. M., World Health Organization, & Boerma, J. T. (2008). *The global burden of disease*. 2004 update (p. 146).

Mathieson, I., & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, *44*(3), 243–246. doi:10.1038/ng.1074

Maynard, T., Haskell, G., Peters, A., Sikich, L., Lieberman, J., & LaMantia, A. (2003).

A comprehensive analysis of 22q11 gene expression in the developing and adult brain. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(24), 14433–14438. doi:10.1073/pnas.2235651100

McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., Dallaire, S., et al. (2006). Common deletion polymorphisms in the human genome. *Nature Genetics*, *38*(1), 86–92.

McCarthy, Shane E, Makarov, V., Kirov, G., Addington, A. M., McClellan, J., Yoon, S., Perkins, D. O., et al. (2009). Microduplications of 16p11.2 are associated with schizophrenia. *Nature Genetics*, *41*(11), 1223–1227. doi:10.1038/ng.474

McGuffin, P., Cohen, S., & Knight, J. (2007). Homing in on Depression Genes. *American Journal of Psychiatry*, *164*(2), 195–197.

McGuffin, P., Katz, R., & Aldrich, J. (1986). Past and present state examination: the assessment of "lifetime ever" psychopathology. *Psychological Medicine*, *16*(2), 461–465.

McGuffin, P., Katz, R., Watkins, S., & Rutherford, J. (1996). A hospital-based twin register of the heritability of DSM-IV unipolar depression. *Archives of General Psychiatry*, *53*(2), 129–136.

McGuffin, P., Knight, J., Breen, G., Brewster, S., Boyd, P. R., Craddock, N., Gill, M., et al. (2005). Whole genome linkage scan of recurrent depressive disorder from the depression network study. *Human Molecular Genetics*, *14*(22), 3337–3345.

McGuffin, P., Rijsdijk, F., Andrew, M., Sham, P., Katz, R., & Cardno, A. (2003). The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Archives of General Psychiatry*, *60*(5), 497–502.

doi:10.1001/archpsyc.60.5.497

McQuillin, A., Bass, N., Anjorin, A., Lawrence, J., Kandaswamy, R., Lydall, G., Moran, J., et al. (2011). Analysis of genetic deletions and duplications in the University College London bipolar disorder case control sample. *European Journal of Human Genetics*, *19*(5), 588–592. doi:10.1038/ejhg.2010.221

Miller, D., Adam, M., Aradhya, S., Biesecker, L., Brothman, A., Carter, N., Church, D., et al. (2010). Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *American Journal of Medical Genetics*, *86*(5), 749–764. doi:10.1016/j.ajhg.2010.04.006

Montcel, Du, S. T., & Mendizabai, H. (1996). Prevalence of 22q11 microdeletion. *Journal of Medical Genetics*, *33*, 719.

Morrow, E. M., Yoo, S.-Y., Flavell, S. W., Kim, T.-K., Lin, Y., Hill, R. S., Mukaddes, N. M., et al. (2008). Identifying autism loci and genes by tracing recent shared ancestry. *Science*, *321*(5886), 218–223. doi:10.1126/science.1157657

Muglia, P., Tozzi, F., Galwey, N. W., Francks, C., Upmanyu, R., Kong, X. Q., Antoniades, A., et al. (2010). Genome-wide association study of recurrent major depressive disorder in two European case-control cohorts. *Molecular Psychiatry*, *15*(6), 589–601. doi:10.1038/mp.2008.131

Mulle, J. G., Dodd, A. F., McGrath, J. A., Wolyniec, P. S., Mitchell, A. A., Shetty, A. C., Sobreira, N. L., et al. (2010). Microdeletions of 3q29 confer high risk for schizophrenia. *American Journal of Medical Genetics*, *87*(2), 229–236. doi:10.1016/j.ajhg.2010.07.013

Murphy, K., Jones, L., & Owen, M. (1999). High rates of schizophrenia in adults

with velo-cardio-facial syndrome. *Archives of General Psychiatry*, *56*(10), 940–945.

Need, A., Ge, D., Weale, M., Maia, J., Feng, S., Heinzen, E., Shianna, K., et al. (2009). A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genetics*, *5*(2), e1000373. doi:10.1371/journal.pgen.1000373

Nielsen, J., & Wohlert, M. (1991). Chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus, Denmark. *Human Genetics*, *87*(1), 81–83.

Niklasson, L., & Gillberg, C. (2010). The neuropsychology of 22q11 deletion syndrome. A neuropsychiatric study of 100 individuals. *Research in Developmental Disabilities*, *31*(1), 185–194. doi:10.1016/j.ridd.2009.09.001

Olshen, A. B., Venkatraman, E. S., Lucito, R., & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, *5*(4), 557–572. doi:10.1093/biostatistics/kxh008

Ou, Z., Berg, J., Yonath, H., & Enciso, V. (2008). Microduplications of 22q11.2 are frequently inherited and are associated with variable phenotypes. *Genetics in Medicine*, *10*(4), 267–277. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18414210

Papolos, D. F., Faedda, G. L., Veit, S., Goldberg, R., Morrow, B., Kucherlapati, R., & Shprintzen, R. J. (1996). Bipolar spectrum disorders in patients diagnosed with velo-cardio-facial syndrome: does a hemizygous deletion of chromosome 22q11 result in bipolar affective disorder? *American Journal of Psychiatry*, *153*(12), 1541–1547.

Parker, M., White, R., Leonardsson, G., Rosewell, I., Ann Jacobs, M., & Milligan, S.

(2000). The nuclear receptor co-repressor Nrip1 (RIP140) is essential for female fertility. *Nature Medicine*, *6*(12), 1368–1374. doi:10.1038/82183

Patten, S. B., Beck, C. A., Williams, J. V. A., Barbui, C., & Metz, L. M. (2003). Major depression in multiple sclerosis: a population-based perspective. *Neurology*, *61*(11), 1524–1527.

Pearson, K. (n.d.). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series 5*. Retrieved March 28, 1900, from http://www.economics.soton.ac.uk/staff/aldrich/1900.pdf

Peiffer, D., Le, J., Steemers, F., Chang, W., Jenniges, T., Garcia, F., Haden, K., et al. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Research*, *16*(9), 1136–1148. doi:10.1101/gr.5402306

Peoples, R., Franke, Y., Wang, Y.-K., Pérez-Jurado, L., Paperna, T., Cisco, M., & Francke, U. (2000). A physical map, including a BAC/PAC clone contig, of the Williams-Beuren syndrome--deletion region at 7q11.23. *American Journal of Medical Genetics*, *66*(1), 47–68. doi:10.1086/302722

Pergadia, M. L., Glowinski, A. L., Wray, N. R., Agrawal, A., Saccone, S. F., Loukola, A., Broms, U., et al. (2011). A 3p26-3p25 genetic linkage finding for DSM-IV major depression in heavy smoking families. *American Journal of Psychiatry*, *168*(8), 848–852. doi:10.1176/appi.ajp.2011.10091319

Perry, G. H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C. W., Scheffer, A., et al. (2008). The Fine-Scale and Complex Architecture of

Human Copy-Number Variation. *American Journal of Medical Genetics*, *82*(3), 685–695. doi:10.1016/j.ajhg.2007.12.010

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, *20*(2), 207–211. doi:10.1038/2524

Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A. C., et al. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology*, *29*(6), 512–520. doi:10.1038/nbt.1852

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, *466*(7304), 368–372. doi:10.1038/nature09146

Polanczyk, G., de Lima, M. S., Horta, B. L., Biederman, J., & Rohde, L. A. (2007). The worldwide prevalence of ADHD: A systematic review and metaregression analysis. *American Journal of Psychiatry*, *164*(6), 942–948. doi:10.1176/appi.ajp.164.6.942

Portnoī, M. (2009). Microduplication 22q11.2: a new chromosomal syndrome. *European Journal of Medical Genetics*, (2-3), 88–93.

Potocki, L., Bi, W., Treadwell-Deering, D., Carvalho, C. M. B., Eifert, A., Friedman, E. M., Glaze, D., et al. (2007). Characterization of Potocki-Lupski syndrome (dup(17)(p11.2p11.2)) and delineation of a dosage-sensitive critical interval that can convey an autism phenotype. *American Journal of Medical Genetics*, *80*(4), 633–649. doi:10.1086/512864

Potter, G. G., Blackwell, A. D., McQuoid, D. R., Payne, M. E., Steffens, D. C., Sahakian, B. J., Welsh-Bohmer, K. A., et al. (2007). Prefrontal white matter lesions and prefrontal task impersistence in depressed and nondepressed elders. *Neuropsychopharmacology*, *32*(10), 2135–2142. doi:10.1038/sj.npp.1301339

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. doi:10.1038/ng1847

Priebe, L., Degenhardt, F. A., Herms, S., Haenisch, B., Mattheisen, M., Nieratschker, V., Weingarten, M., et al. (2011). Genome-wide survey implicates the influence of copy number variants (CNVs) in the development of early-onset bipolar disorder. *Molecular Psychiatry*, 1–12. doi:10.1038/mp.2011.8

Pruitt, K. D. (2004). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *33*(Database issue), D501–D504. doi:10.1093/nar/gki025

Pruitt, K. D., Tatusova, T., Klimke, W., & Maglott, D. R. (2009). NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research*, *37*(Database), D32–D36. doi:10.1093/nar/gkn721

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Medical Genetics*, *81*(3), 559–575. doi:10.1086/519795

Quintero-Rivera, F., Sharifi-Hannauer, P., & Martinez-Agosto, J. A. (2010). Autistic and psychiatric findings associated with the 3q29 microdeletion syndrome:

case report and review. *American Journal of Medical Genetics*, *152A*(10), 2459–2467. doi:10.1002/ajmg.a.33573

R Development Core Team. (2005). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

Ratcliffe SG, Tierney I, Nshaho J, Smith L, Springbett A, & S, C. (1982). Edinburgh study of growth and development of children with sex chromosomal abnormalities. (S. DA, Ed.)*Birth Defects*, *18*, 41–60.

Redon, R., Ishikawa, S., Fitch, K., Feuk, L., Perry, G., Andrews, T., Fiegler, H., et al. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. doi:10.1038/nature05329

Rietschel, M., Mattheisen, M., Frank, J., Treutlein, J., Degenhardt, F., Breuer, R., Steffens, M., et al. (2010). Genome-wide association-, replication-, and neuroimaging study implicates HOMER1 in the etiology of major depression. *Biological Psychiatry*, *68*(6), 578–585. doi:10.1016/j.biopsych.2010.05.038

Ripke, S., Sanders, A. R., Kendler, K. S., Levinson, D. F., Sklar, P., Holmans, P. A., Lin, D.-Y., et al. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature Publishing Group*, *43*(10), 969–976. doi:10.1038/ng.940

Risch, N., & Merikangas, K. (1996). The Future of Genetic Studies of Complex Human Diseases. *Science*, *273*(5281), 1516–1517. doi:10.1126/science.273.5281.1516

Robin, N., & Shprintzen, R. (2005). Defining the clinical spectrum of deletion 22q11.2. *Journal of Pediatrics*, *147*(1), 90–96. doi:10.1016/j.jpeds.2005.03.007

Rodriguez-Santiago, B., Malats, N., Rothman, N., Armengol, L., Garcia-Closas, M., Kogevinas, M., Villa, O., et al. (2010). Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. *American Journal of Medical Genetics*, *87*(1), 129–138. doi:10.1016/j.ajhg.2010.06.002

Rucker, J. J. H., Breen, G., Pinto, D., Pedroso, I., Lewis, C. M., Cohen-Woods, S., Uher, R., et al. (2011). Genome-wide association analysis of copy number variation in recurrent depressive disorder. *Molecular Psychiatry*, –. doi:10.1038/mp.2011.144

Rujescu, D., Ingason, A., Cichon, S., Pietiläinen, O. P. H., Barnes, M. R., Toulopoulou, T., Picchioni, M., et al. (2009). Disruption of the neurexin 1 gene is associated with schizophrenia. *Human Molecular Genetics*, *18*(5), 988–996. doi:10.1093/hmg/ddn351

Sahay, A., Drew, M., & Hen, R. (2007). Dentate gyrus neurogenesis and depression. *Progress in Brain Research*, *163*, 697–722.

Saitta, S. C., Harris, S. E., Gaeth, A. P., Driscoll, D. A., McDonald-McGinn, D. M., Maisenbacher, M. K., Yersak, J. M., et al. (2004). Aberrant interchromosomal exchanges are the predominant cause of the 22q11.2 deletion. *Human Molecular Genetics*, *13*(4), 417–428. doi:10.1093/hmg/ddh041

Schena, M., Shalon, D., Davis, R., & Brown, P. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*(5235), 467–470.

Scherer, S., Lee, C., Birney, E., Altshuler, D., Eichler, E., Carter, N., Hurles, M., et al. (2007). Challenges and standards in integrating surveys of structural variation. *Nature Genetics*, *39*(7 Suppl), S7–15. doi:10.1038/ng2093

Schulze, T. G., Detera-Wadleigh, S. D., Akula, N., Gupta, A., Kassem, L., Steele, J., Pearl, J., et al. (2009). Two variants in Ankyrin 3 (ANK3) are independent genetic risk factors for bipolar disorder. *Molecular Psychiatry*, *14*(5), 487–491. doi:10.1038/mp.2008.134

Schwab, S. G., & Wildenauer, D. B. (2009). Update on key previously proposed candidate genes for schizophrenia. *Current Opinion in Psychiatry*, *22*(2), 147–153. doi:10.1097/YCO.0b013e328325a598

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., et al. (2007). Strong association of de novo copy number mutations with autism. *Science*, *316*(5823), 445–449. doi:10.1126/science.1138659

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., et al. (2004). Large-scale copy number polymorphism in the human genome. *Science*, *305*(5683), 525–528. doi:10.1126/science.1098918

Sen, A., & Srivastava, M. S. (1975). On Tests for Detecting Change in Mean. *The Annals of Statistics*, *3*(1), 98–108. doi:10.1214/aos/1176343001

Shaikh, T. H., Kurahashi, H., Saitta, S. C., O'Hare, A. M., Hu, P., Roe, B. A., Driscoll, D. A., et al. (2000). Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Human Molecular Genetics*, *9*(4), 489–501.

Sharp, A., Mefford, H., Li, K., Baker, C., Skinner, C., Stevenson, R., Schroer, R., et al. (2008). A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genetics*, *40*(3), 322–328. doi:10.1038/ng.93

Shi, J., Potash, J. B., Knowles, J. A., Weissman, M. M., Coryell, W., Scheftner, W. A.,

Lawson, W. B., et al. (2011). Genome-wide association study of recurrent early-onset major depressive disorder. *Molecular Psychiatry*, *16*(2), 193–201. doi:10.1038/mp.2009.124

Shinawi, M., Schaaf, C. P., Bhatt, S. S., Xia, Z., Patel, A., Cheung, S. W., Lanpher, B., et al. (2009). A small recurrent deletion within 15q13.3 is associated with a range of neurodevelopmental phenotypes. *Nature Genetics*, *41*(12), 1269–1271. doi:10.1038/ng.481

Shprintzen, R. J. (2001). *Management of genetic syndromes*. (S. B. Cassidy & J. Allanson, Eds.)Velo-cardio-facial syndrome (1st ed. pp. 615–632). New York: Wiley.

Shprintzen, R. J., Goldberg, R. B., Lewin, M. L., Sidoti, E. J., Berkman, M. D., Argamaso, R. V., & Young, D. (1978). A new syndrome involving cleft palate, cardiac anomalies, typical facies, and learning disabilities: velo-cardio-facial syndrome. *The Cleft Palate Journal*, *15*(1), 56–62.

Shyn, S., Shi, J., Kraft, J., Potash, J., Knowles, J., Weissman, M., Garriock, H., et al. (2009). Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. *Molecular Psychiatry*, *16*, 202–215. doi:10.1038/mp.2009.125

Sklar, P., Ripke, S., Scott, L. J., Andreassen, O. A., Cichon, S., Craddock, N., Edenberg, H. J., et al. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics*, *43*(10), 977–983. doi:10.1038/ng.943

Sklar, P., Smoller, J. W., Fan, J., Ferreira, M. A. R., Perlis, R. H., Chambert, K.,

Nimgaonkar, V. L., et al. (2008). Whole-genome association study of bipolar disorder. *Molecular Psychiatry*, *13*(6), 558–569. doi:10.1038/sj.mp.4002151

Southern, E. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, *98*(3), 503–517.

Southern, E. (2006). Southern blotting. *Nature Protocols*, *1*(2), 518–525. doi:10.1038/nprot.2006.73

St Clair, D., Blackwood, D., Muir, W., Walker, M., St Clair, D., Muir, W., Carothers, A., et al. (1990). Association within a family of a balanced autosomal translocation with major mental illness. *The Lancet*, *336*(8706), 13–16. doi:10.1016/0140-6736(90)91520-K

Steemers, F., Chang, W., Lee, G., Barker, D., Shen, R., & Gunderson, K. (2006). Whole-genome genotyping with the single-base extension assay. *Nature Methods*, *3*(1), 31–33. doi:10.1038/nmeth842

Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, O., Ingason, A., Steinberg, S., Fossdal, R., et al. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature*, *455*(7210), 232–236. doi:10.1038/nature07229

Sudhof, T. (2008). Neuroligins and neurexins link synaptic function to cognitive disease. *Nature*, *455*(7215), 903–911. doi:10.1038/nature07456

Sullivan, P. F. (2005). The genetics of schizophrenia. *PLoS Medicine*, *2*(7), e212. doi:10.1371/journal.pmed.0020212

Sullivan, P. F., de Geus, E. J. C., Willemsen, G., James, M. R., Smit, J. H., Zandbelt, T., Arolt, V., et al. (2009). Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Molecular*

Psychiatry, *14*(4), 359–375. doi:10.1038/mp.2008.125

Sullivan, P. F., Neale, M. C., & Kendler, K. S. (2000). Genetic epidemiology of major depression: review and meta-analysis. *American Journal of Psychiatry*, *157*(10), 1552–1562.

Szatmari, P., Paterson, A., Zwaigenbaum, L., Roberts, W., Brian, J., Liu, X., Vincent, J., et al. (2007). Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature Genetics*, *39*(3), 319–328. doi:10.1038/ng1985

Terracciano, A., Sanna, S., Uda, M., Deiana, B., Usala, G., Busonero, F., Maschio, A., et al. (2010). Genome-wide association scan for five major dimensions of personality. *Molecular Psychiatry*, *15*(6), 647–656. doi:10.1038/mp.2008.113

Thomson, W., Barton, A., Ke, X., Eyre, S., Hinks, A., Bowes, J., Donn, R., et al. (2007). Rheumatoid arthritis association at 6q23. *Nature Genetics*, *39*(12), 1431–1433. doi:10.1038/ng.2007.32

Tsuang, D. W., Millard, S. P., Ely, B., Chi, P., Wang, K., Raskind, W. H., Kim, S., et al. (2010). The effect of algorithms on copy number variant detection. *PLoS One*, *5*(12), e14456–. doi:10.1371/journal.pone.0014456

Tsuang, M. T., Winokur, G., & Crowe, R. R. (1980). Morbidity risks of schizophrenia and affective disorders among first degree relatives of patients with schizophrenia, mania, depression and surgical conditions. *The British Journal of Psychiatry*, *137*, 497–504.

Tsuang, M., & Faraone, S. (1990). *The genetics of mood disorders* (1st ed.). The Johns Hopkins University Press.

Uher, R. (2009). The role of genetic variation in the causation of mental illness:

an evolution-informed framework. *Molecular Psychiatry*, *14*, 1072–1082. doi:10.1038/mp.2009.85

Uher, R., Maier, W., Hauser, J., Marusic, A., Schmael, C., Mors, O., Henigsberg, N., et al. (2009). Differential efficacy of escitalopram and nortriptyline on dimensional measures of depression. *British Journal of Psychiatry*, *194*(3), 252–259. doi:10.1192/bjp.bp.108.057554

Valero, R., Marfany, G., González-Angulo, O., González-González, G., Puelles, L., & Gonzàlez-Duarte, R. (1999). USP25, a novel gene encoding a deubiquitinating enzyme, is located in the gene-poor region 21q11.2. *Genomics*, *62*(3), 395–405. doi:10.1006/geno.1999.6025

van de Laar, I., Rabelink, G., Hochstenbach, R., Tuerlings, J., Hoogeboom, J., & Giltay, J. (2002). Diploid/triploid mosaicism in dysmorphic patients. *Clinical Genetics*, *62*(5), 376–382.

Van der Aa, N., Rooms, L., Vandeweyer, G., van den Ende, J., Reyniers, E., Fichera, M., Romano, C., et al. (2009). Fourteen new cases contribute to the characterization of the 7q11.23 microduplication syndrome. *European Journal of Human Genetics*, *52*(2-3), 94–100. doi:10.1016/j.ejmg.2009.02.006

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., et al. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351. doi:10.1126/science.1058040

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269. doi:10.1109/TIT.1967.1054010

Vorstman, J. A. S., Staal, W. G., van Daalen, E., Van Engeland, H., Hochstenbach, P.

F. R., & Franke, L. (2005). Identification of novel autism candidate regions through analysis of reported cytogenetic abnormalities associated with autism. *Molecular Psychiatry*, *11*(1), 18–28. doi:10.1038/sj.mp.4001757

Walsh, T., McClellan, J., McCarthy, S., Addington, A., Pierce, S., Cooper, G., Nord, A., et al. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, *320*(5875), 539–543. doi:10.1126/science.1155174

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S., Hakonarson, H., et al. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, *17*(11), 1665–1674. doi:10.1101/gr.6861907

Wang, Y., Chen, J., Chen, M., Chen, C., Lai, I., Chen, T., Hong, C., et al. (2008). Neuregulin 3 genetic variations and susceptibility to schizophrenia in a Chinese population. *Biological Psychiatry*, *64*(12), 1093–1096. doi:10.1016/j.biopsych.2008.07.012

Watanabe, K., Yamada, H., & Yamaguchi, Y. (1995). K-glypican: a novel GPI-anchored heparan sulfate proteoglycan that is highly expressed in developing brain and kidney. *Journal of Cell Biology*, *130*(5), 1207–1218.

Watson, J. D., & Crick, F. H. C. (1953, April 25). Molecular Structure of Nucleic Acids. A Stucture for Deoxyribose Nucleic Acid. *Nature*. Retrieved January 13, 2012, from http://www.nature.com/nature/dna50/watsoncrick.pdf

Weiss, L. A., Shen, Y., Korn, J. M., Arking, D. E., Miller, D. T., Fossdal, R., Saemundsen, E., et al. (2008). Association between Microdeletion and

Microduplication at 16p11.2 and Autism. *New England Journal of Medicine*, *358*(7), 667–675. doi:10.1056/NEJMoa075974

Weissman, M. M., Gershon, E. S., Kidd, K. K., Prusoff, B. A., Leckman, J. F., Dibble, E., Hamovit, J., et al. (1984). Psychiatric Disorders in the Relatives of Probands With Affective Disorders: The Yale University--National Institute of Mental Health Collaborative Study. *Archives of General Psychiatry*, *41*(1), 13. doi:10.1001/archpsyc.1984.01790120015003

Weissman, M. M., Leaf, P. J., Tischler, G. L., Blazer, D. G., Karno, M., Bruce, M. L., & Florio, L. P. (1988). Affective disorders in five United States communities. *Psychological Medicine*, *18*(1), 141–153.

Weissman, M. M., Wickramaratne, P., Adams, P. B., Lish, J. D., Horwath, E., Charney, D., Woods, S. W., et al. (1993). The relationship between panic disorder and major depression. A new family study. *Archives of General Psychiatry*, *50*(10), 767–780.

Wender, P. H., Kety, S. S., Rosenthal, D., Schulsinger, F., Ortmann, J., & Lunde, I. (1986). Psychiatric Disorders in the Biological and Adoptive Families of Adopted Individuals With Affective Disorders. *Archives of General Psychiatry*, *43*(10), 923. doi:10.1001/archpsyc.1986.01800100013003

Williams, N. M., Zaharieva, I., Martin, A., Langley, K., Mantripragada, K., Fossdal, R., Stefansson, H., et al. (2010). Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. *The Lancet*, *376*(9750), 1401–1408. doi:10.1016/S0140-6736(10)61109-9

Winchester, L., Yau, C., & Ragoussis, J. (2009). Comparing CNV detection methods for SNP arrays. *Briefings in Functional Genomics and Proteomics*, *8*(5), 353–

366. doi:10.1093/bfgp/elp017

Wing, J., Babor, T., Brugha, T., Burke, J., Cooper, J., Giel, R., Jablenski, A., et al. (1990). SCAN. Schedules for Clinical Assessment in Neuropsychiatry. *Archives of General Psychiatry*, *47*(6), 589–593.

World Health Organisation. (1992). *The ICD-10 Classification of Mental and Behavioural Disorders : Clinical Descriptions and Diagnostic Guidelines*. World Health Organization.

World Health Organisation. (1993). *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. World Health Organization.

World Health Organisation. (n.d.). World Health Organisation | Depression. *WHO*. World Health Organization. Retrieved November 8, 2011, from http://www.who.int/mental_health/management/depression/definition/en/

Wray, N. R., Pergadia, M. L., Blackwood, D. H. R., Penninx, B. W. J. H., Gordon, S. D., Nyholt, D. R., Ripke, S., et al. (2012). Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Molecular Psychiatry*, *17*(1), 36–48. doi:10.1038/mp.2010.109

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, *14*(4), 1261–1295.

Xu, B., Roos, J. L., Levy, S., van Rensburg, E. J., Gogos, J. A., & Karayiorgou, M. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. *Nature Genetics*, *40*(7), 880–885. doi:10.1038/ng.162

Xu, B., Woodroffe, A., Rodriguez-Murillo, L., Roos, J., van Rensburg, E., Abecasis, G.,

Gogos, J., et al. (2009). Elucidating the genetic architecture of familial schizophrenia using rare copy number variant and linkage scans. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(39), 16746–16751. doi:10.1073/pnas.0908584106

Yobb, T. M., Somerville, M. J., Willatt, L., Firth, H. V., Harrison, K., MacKenzie, J., Gallo, N., et al. (2005). Microduplication and triplication of 22q11.2: a highly variable syndrome. *American Journal of Medical Genetics*, *76*(5), 865–876. doi:10.1086/429841

Yu, Y., & Williams, D. R. (1999). Handbook of the sociology of mental health. In C. S. Aneshensel & J. C. Phelan (Eds.), *Socioeconomic Status and Mental Health* (1st ed.). New York: Kluwer Academic/Plenum.

Zeggini, E. (2007). A new era for Type 2 diabetes genetics. *Diabetic Medicine*, *24*(11), 1181–1186. doi:10.1111/j.1464-5491.2007.02274.x

Zhang, D., Cheng, L., Qian, Y., Alliey-Rodriguez, N., Kelsoe, J., Greenwood, T., Nievergelt, C., et al. (2009a). Singleton deletions throughout the genome increase risk of bipolar disorder. *Molecular Psychiatry*, *14*(4), 376–380. doi:10.1038/mp.2008.144

Zhang, F., Gu, W., Hurles, M., & Lupski, J. (2009b). Copy number variation in human health, disease, and evolution. *Annual Reviews in Genomics and Human Genetics*, *10*, 451–481. doi:10.1146/annurev.genom.9.081307.164217

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, *18*(6), 292–298. doi:10.1016/S0169-5347(03)00033-8

Zis, A. P., & Goodwin, F. K. (1979). Major affective disorder as a recurrent illness:

a critical review. *Archives of General Psychiatry*, *36*(8 Spec No), 835–839.